Mathematical Statistics
Stockholm University

Customer duration in non-life insurance industry

Erik Gustafsson

Examensarbete 2009:3

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2009:3**,
http://www.math.su.se/matstat

Customer duration in non-life insurance industry

Erik Gustafsson*

April 2009

Abstract

The subject of this master thesis is the application of survival analysis to predict policy churns in a non-life insurance industry. Especially, models and methods are applied to estimate survival probabilities on customer-level in a competing risk setting, where churns occur of different types of causes. By following motor policy holders over a 3-year period, probabilites are estimated which enables scoring of customers, especially those likely to churn within this time period. Cause-specific semiparametric hazard functions are modelled with Cox regression given customer data at the beginning of the study period. The models are estimated from data on private customers in the Danish insurance company Codan. The main conclusion is that time-fixed covariate and time-invariant effect models that are used for prediction here might be an over-simplification of churns on customer-level, as they disregard the impact of customers-specific events during follow-up. This suggests more flexible models when analysing churns.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: erik.gustafsson@codanforsikring.no Supervisor: Joanna Tyrcha

Customer duration in non-life insurance industry

Erik Gustafsson

Abstract

The subject of this master thesis is the application of survival analysis to predict policy churns in a non-life insurance industry. Especially, models and methods are applied to estimate survival probabilities on customer-level in a competing risk setting, where churns occur of different types of causes.

By following motor policy holders over a 3-year period, probabilites are estimated which enables scoring of customers, especially those likely to churn within this time period.

Cause-specific semiparametric hazard functions are modelled with Cox regression given customer data at the beginning of the study period. The models are estimated from data on private customers in the Danish insurance company Codan.

The main conclusion is that time-fixed covariate and time-invariant effect models that are used for prediction here might be an over-simplification of churns on customer-level, as they disregard the impact of customers-specific events during follow-up. This suggests more flexible models when analysing churns.

Contents

Acknowledgements

I would like to give my thanks to those who have helped me and supported me during my work with this master thesis. Especially, I would like to thank my supervisor at Stockholm University, Joanna Tyrcha, for useful comments along the way and patience.

I would also like to thank my colleagues at Customer & Channel Development in Trygg-Hansa in Stockholm, Sweden and Codan in Copenhagen, Denmark for support and discussions. In particular, Paul Andersson, for introducing me to the subject and Eva Sandberg, for expertice. A special thanks to Jim Gustafsson at Codan in Denmark for introducing me to the insurance industry in the first place.

Finally, my sincere thanks to my new colleagues at Codan Forsikring in Oslo, Norway for a warm welcome and professor Ørnulf Borgan at Oslo University for shown interest, valuable comments on my work and theoretical discussions.

1 Introduction

Many (non-life) insurance markets are saturated and competitive ones, where customers almost instantaneously can cancel insurance policies, according to statute. Especially, this enables customers to transfer insurances to a competeting insurance company, a process where cancellation is notified by either the customer or the competing company. Respective of which, there will be little or no time trying to retain the customer.

This has resulted in an increased interest in customer relationship management (CRM) over the last decades, in which relationships are monitored and analyzed. Since insurance policy holders are associated with risks - the risk of having future claims, delayed payments etc. - but also with needs, companies want to learn as much as possible of these factors in order to conclude when focus should be put on a customer. And since each customer show case its own set of risks and needs, it is crucial for this relationship management to be specific and accurate in its conclusions.

Somewhat simplified, successful CRM results in customer satisfaction, which is substantial for customer longevity.

There is, however, an outflow of customers in an insurance company. This can be seen as a transition from a state of being an active customer to a state of being inactive. Especially, this transistion could be due to different causes, as seen in fig. 1, since customers cancel their policies for different causes.

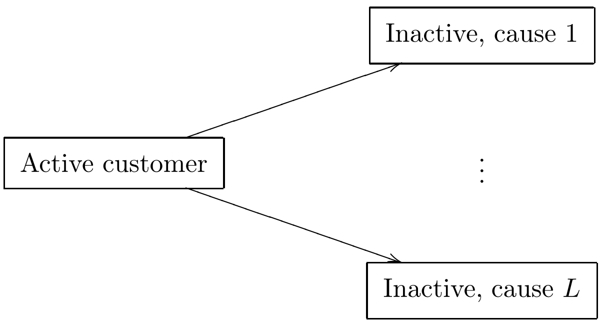


Figure 1: *Outflow of customers. Each customer can cancel an insurance of one out of L different causes.*

There is also an inflow: new customers are attracted by marketing, campaigns and other means. Due to the saturated and mature nature of insurance markets, where companies offer similar products and services and most potential customers are most likely to be found in competeting companies, attracting customers that already have insurance coverage from a competitor is of great importance. Lastly, a substantial part of the customers will renew their policies, extending their customer duration and displaying their attitude towards the company. Most commenly, non-life insurance policies run for one year and are renewed unless notice of cancellation has been given, in which case the cancellation will have its impact from the written policy's date of expiration.[1]

Insurance companies, like other types of customer based businesses, wish to optimize their customer base by controlling the in- and outflow, as described above, of customers. Different strategies to accomplish this are available, but common for all is the need of knowledge of market conditions and customer behaviours. When deducing underlying principles of the latter, especially on customer longevity, analysis of cancellations is necessary.

This will give the company an idea on when the risk of losing a customer is high/low and what causes this. It also gives the ability to identify high-risk customers, which is useful when targeting new customers in campaigns, since it enables customer segmentation highlighting customers with high/low predicted lifetime duration.

Of course, longevity is not all when it comes to the profitability of a customer, but should be looked at with respect to premiums and margins in order to seperate high profitable customers from other customers. This produces a powerful tool for customer segmentation and enables the company to further direct and optimize campaigns, by using this customer value.

[1] For different countries and markets, different rules apply as to when a note of cancellation has its effect. For the Danish private market, which is considered here, the cancellation will have its impact on the renewal date.

2 Preliminaries

This section gives the definitions and theoretical concepts that are used throughout the text. It starts by reviewing the purpose of the study as well as its applications in 2.1, along with some notes on previous work in related fields in 2.2. In 2.3 and 2.4, the necessary non-technical concepts are definied and discussed. Finally, in 2.5, some of the basic technical concepts of survival analysis are reviewed.

2.1 Objectivies

In this thesis we are concerned with customer tenure, more specificely on customer duration, in a non-life insurance company. The main purpose is to model duration in order to score (i.e. rank) customers, thus be able to answer the question:

Which customer is more likely to stay?

This is of course a very general question, and will be further specified in order to be answered.

By applying statistical methods from survival analysis, we suggest methods to estimate customer survival function and customer hazard function on product-level in order to gain insight in customer behaviours. This will be used to develop time until event regression models, more specifically Cox models (Cox [7]), including and measuring effects relevant for duration.

The analysis will consider competing risks, i.e. we will investigate duration in the presence of different cancellation causes. Using these cause-specific effects, we develop fixed covariate and time invariant effect models to predict survival probabilities over a finite time period, given initial covariates. This is done by stacking data and performing stratified analysis and allowing for cause-specific effects to be equal over strata.

In practice, the time period could be set to 12 month, in which case the survival probability would be intepreted as the probability of renewal.

The reason for including competing risks is to conclude whether or not it is relevant and possible to pinpoint which covariates effect which causes. This will be useful when focusing on retention and other CRM activities, since it enables scoring of customers that are more likely to cancel their policies from causes of certain interest for the company.

We will also look at the possibility to include time-varying effects, thus fitting a generalized Cox model to the data, as suggested in e.g. Therneau and Grambsch [21]. In this part will seek to find effects that might vary over time and thus violates some assumptions on time-invarians that are imposed under certain duration models.

The model development is preceded by a fairly thoroughly presentation of basic concepts and topics in survival analysis. The development part will be based on and presented in the framework of martingale theory. Specifically, the models are based on the theory of counting processes.

Furthermore, we will profile high respectively low loyalty customer, thus highlighting customer characteristics which are useful for market targeting.

The model is applied to a data set from the Danish non-life insurance company Codan, part of the RSA Group.

2.2 Previous studies

Overall, very little has been written on survival analysis and customer lifetime duration applied to the insurance industry. Other industries are more well-represented. In Lu [16], churn analysis is carried out in a telecommunication company by looking at historical customer data which is used for predictive modelling of customer duration.

In van den Poel and Larivère [22], a similar model is used with time-varying data when predicting churn incidences in the financial service market.

Other industries where survival analysis has been proved useful are marketing science, political science and reliability engineering.

Guillen, Nielsen, Scheike and Perez-Marin [11], however, analyze customer lifetime duration in the insurance industry and apply an extended Cox model to retention time after an initial, partial cancellation of insurance policies. They find empirical evidence of time-dependent effects of factors explaining duration and suggest methods to identify customers with high risk of cancelling all remaining policies and how the risk varies over time.

Our setting is somewhat different, in that we will develop models with respect to products - i.e. customers will be scored for each product (listed below) that is considered here - in the presence of competing risks.

2.3 Definitions and Exclusions

Here, a few concepts from customer relationship management (CRM) and the insurance industry are defined, which will appear, or already have appeared, in the text.

Customer - in this context a domestic household, containing one or more policy holders, represented by a primary policy holder with at least one active insurance at the beginning of the study period. The primary customer is either a person or a company, small enough to, in a sense, be considered as a person.[2]

Focus Customer (FC) - customers who have an insurance portfolio containing at least one of the following insurance policies: home, car, accident and comprehensive household. These policies are refereed to as focus products (FP).[3]

Churn - a broad term of the action of cancelling an service, which in this case translates to cancelling insurance coverage. Furthermore, this cancellation

[2] These customers are those who will be used in campaigns and thus will be scored.

[3] We exclude customers holding life and pension insurance policies, carried on from the time when the company held this type of policies.

must not be disrupted by renewal[4] within 12 months. In other words, a customer is not considered churned if the customer returns within one year, i.e. we consider *net churns*.

Loyalty - here used in the sense of longevity, i.e. customer duration measured in some fixed time unit. Thus, a loyal customer is someone who will stay within the company for a considerable amount of time intervals.

Profitability - with respect to the concept of focus products, all focus customers are considered profitable, since these products correspond to the greatest part of the premium volume and/or are given most attention in direct sale marketing and campaigns. We disregard the cost of claims in this definition.

The necessity of these definitions of customer-centric metrics are not only because of precision in statistical assumptions and interpretation of results, but for business metrics in general; as Berry and Linoff [6] remark, a business often becomes what it is measured by.

Some clearifications of these concepts might be needed. Since our main objective is to model customer duration, we are interested in the time until the relationship with the customer has ended in the sense of net churns. That is, we are not really interested in the object being covered by the insurance policy. For example, assume a car assurer sells the car which is covered by a insurance company. The need of coverage ceases, since there is no longer any risk, or object, to insure. The relationship, however, does not necessarily end with this: there might still be another person in the household with a car insurance.

When it comes to the actual churning occurence, the *churning point*, we let this be defined as the date at which the policy is deactivated. This date is not always possible to detect in other businesses, where churns are not open, such as various product subscriptions businesses.

Loyalty, or rather the causal relationship between loyalty and its ancedents, is often described by behavioural and (relative) attitudinal factors (Dick and Basu [9] and Beerli, Martín and Quintana [5]). E.g. how is a customer's attitude towards the insurance company effected by a shift in premiums or how do customers behave in case of inefficient claim handling processes? Although this is, conceptually, a more appealling view of loyalty compared to customer duration it is not as obvious how to measure its effects over time. Especially, it is rather difficult to see the direct effect of a specific CRM activity on loyalty since, during the follow-up time of this activity, the customers probably will undergo other activites (e.g. marketing, claims handling, information) which also will effect their attitude towards the company. The problem here is that it takes time to observe time and conditions are not constant during the course of time.

Hence, one might argue that the concept of loyalty is not fully utilized here in favor of measurability. Thus, "disloyalty" is synonymous with "churning" in that

[4]This 'final' cancellation of insurance coverage is known as *lapsing* in some parts of the world, where churn would imply cancellation by replacement of another policy.

sense.

Furthermore, in the definition of a customer we exclude large companies and instead focus on private households and smaller companies, that are subject to conventional products through campaigns.

In terms of these definitions, our objective is to estimate customer duration for focus customers and doing this for each focus product - i.e. one estimated probability for each type of focus product - by investigating churns over time.

2.4 Classification of risks

As mentioned above, our analysis considers competing risks; customers can churn for different causes. Typically, risk and cause refer to the same condition, but the reference differ with respect to the occurrence of the event. In other words, a "risk" becomes a "cause" if an event has occurred. This explains why it is difficult to measure loyalty in the sense of attitude towards the company, since "circumstances" - which affects the attitude - and "cause" do not necessarily have this relation.

However, "risk" and "cause" are affected and explained by attitude and circumstances respectively, but not necessarily jointly. E.g. if a customer chooses to churn, this decision is based on attitude - which in turn is based on previous happenings - but this attitude is not the cause of churning: it is the circumstance *yielding* the cause of churning and is refered to as risk.

Here we classify the competing risks used in the study[5]:

I. Insurance policy transferred to competitor

II. Insurance policy cancelled on internal initiative

III. Need of coverage ceased

IV. Insurance policy cancelled by customer

V. Residual

Examples of I. would be when a customer moves the primary policy to a competing insurance company (which could be because of e.g. price or terms).

Case II. occurs when the company has remarks on a customer, e.g. if the customer has delayed payments or is considered as too risky with respect to underwriting criterias, and chooses to cancel the policy according to legislation.

Case III. consists of those not having any need of insurance, e.g. if the insured house is sold and the assured moves to an apartment or customer has deceased.

Case IV. is similiar to I. in that they are both (more or less) voluntary and initiated by the customer, but in the former case the company does not know if the risk will be covered by another company. In fact, the risk could be covered by another internal insurance held by a policy holder within the same household, in which case the relation may or may not be retained. An example of this type would churning due to adjustments in premiums or due to household reconstellations.

[5]Originally there were 32 risks, internally defined by the insurance company

Finally, there is a residual group, consisting of churns that can be seen as mixtures of cases I-IV, policies cancellations that are not as easily grouped as the others. Also, this group has showed to be small in numbers, compared to the other groups.

Overall, this classification is useful since it greatly reduces the causes of churning, to a more suitable number, upon which the analysis will be performed. In fact, there might be necessary to further regroup causes when some of the groups I.-V. are too rare to make inference, in which case one might draw suspect conclusions about certain covariates effect on these rare events. When this is the case, the regrouping will typically depend on which product we are analyzing, but we will usually link I. with IV. and II. with V., leaving III. unchanged. This reclassification is then somewhat similar to one common in other industries where churning is treathed as either voluntary, unvoluntary or expected, see Berry and Linoff [6].

The main purpose of investigating these different types of risk is to conclude whether or not customers belonging to different risk groups have different survival times. Also, different groups are interesting to the company for different reasons: II. and IV. tells us something about customer behaviours and the subject being insured, whereas I. might tell us something about market conditions and indicate that a competitor offers a lower price or different terms.

Due to our setting it is possible to belong to different risk-classes at the same time. In order to perform unique classification, we set up rules for this situations. This is of course only a problem when several insurances - all of the same focus product - are churned on the same date. Otherwise, we only look at the last date of churns and simply pick that reason as the reason of why the relation has ended. For ties we let customer belong to I., if this class is represented. If not, we choose class IV., otherwise III and then II.

This way, we favor churns that are initiated by the customer and for some situations this might be appropriate, for other not. See section 6.1 for comments on this and other data related issues.

To illustrate the classification (and the issue with statistical power for some of the classes) consider the table below, which gives the number of churned customers due to each cause during the study period. Here class 0 is of course those customers who didn't churn at all.

As seen, more than 40 % of those customers holding at least one motor policy at the beginning of 2005, churned within three years. Among these customers, the most common cause of churning was that of transferring the policy to a competing company, followed by those due to ceased risk. The remaining risk classes make up for 5 % of the churns.

2.5 Concepts from Survival Analysis

2.5.1 Essentials

Let T represent the time measured in some unit for an object from some reference point, called *time of origin*, up to the occurence of an event or to the end of the

Class	# of churns	churns, %
0.	48605	57.78
I.	22321	26.54
II.	693	0.82
III.	11248	13.37
IV.	726	0.86
V.	521	0.62
I.-V.	35509	42.22

Table 1: *Churns during 2005-2007 for motor insurance.*

study period. The lenght of this interval is the survival time and is treated as a non-negative continuous random variable with a cumulative distribution function $F_T(t) = P(T \leq t)$ and probability density funtion $f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t}F_T(t)$. (Note that $T \in (0, \infty)$, but in some cases it may be more realistic to let T be finite.)[6]

Along with the survival time is the event, which either occurs or does not occur during the study period. Examples of an event could be death, divorce or some medical diagnos. For this reason introduce a dummy status variable, δ, to indicate whether the event of the study has occured or not, i.e.

$$\delta = \begin{cases} 1, \text{if event has occured} \\ 0, \text{if event has not occured} \end{cases}$$

In our case, an event has occured, i.e. $\delta = 1$, once all of the insurances within a focus product have been cancelled. As we defined risk classes in 2.4, we realize that an event could occur due to different reasons and we will come back to this in 2.5.4.

A third component is of course explanatory information on objects, which is given at the start of the study and recorded during the study period.

2.5.2 Censored and truncated data

Survival data is usually subject to various censoring schemes and other special features, which will be mentioned here. For more details, see Klein and Moeschberger [15].

What is known about a subject, is that eventually the event will *de facto* occur or already has occured, but not when. The must common case is the former - where subjects have not had the event during study period, but might have it directly after - and is known as *right censoring*. It is also possible that subjects drop out of the study, for whatever reason, making data incomplete. Due to the construction of insurance policies, the date of churn will always be known and, hence, censored observation are those who haven't had an event at the end of the study period. In other words, all censored observations are so, simply because we have not followed

[6] This cases could then be dealt with by using conditional survival and hazard functions.

them long enough. Due to the (right) censoring, no insight can be obtained about the upper tail of the ditstribution of the lifetime.

There are, however, some complications in this matter in a competing risks setting: subjects who have had an event are considered censored from risks other than that which caused the event. Alternatively the risks could be assumed to be independent, i.e. the occurence of an event due to some cause would not effect the likelihood for an event due to some other, had it not happen. See e.g. Crowder [8] for more discussion on this theme.

If the event is only known to have occured before some point in time, this is known as *left censoring*. If the time of event occurrence is known to be somewhere between two times, this is known as *interval censoring*. Right and left censoring can thus be seen as special cases of interval censoring, with right time limit set to ∞ and left time limit set to 0.

Another feature in survival data is *truncation*. *Left truncation* is present if subjects have been exposed to the risk of having the event before participating in the study, e.g. if a customer had a focus product before time of origin. *Right* truncation is nothing else than left censoring.

All of the above mentioned censoring schemes can emerge from different study designs. If the subjects in the study are followed for a fixed, predetermined period of time, they are subject to *type I* censoring. Here, the number of events that may occur is random, but the study period is fixed. A generalization of this, called *fixed censoring*, would be to let each subject have its own potential maximum observation time, τ_i.

If they instead are followed until a fixed (in advance) number of the subjects, d, have had the event, they are subject to *type II* censoring. In this case, study period is random and d determines the precision of the study and could be used as a design parameter. In a more general setting, called *random censoring*, each subject is associated with a potential censoring time, C_i, and a potential lifetime, T_i.In this design, observations are terminated for reasons that are not under control. It also is possible that entry times vary randomly across subjects and these entry times are not decided by the study. The entry times for customers in an insurance company could typically be considered as random.

When dealing with survival time, one must consider what time variabel is observed, i.e. if it is time until event, T, or time until censoring, C. For right censored data, we will observe $\widetilde{T}_i = \min(T_i, C_i)$ for subject i. The censoring time is also considered as random, with the same restrictions as T and assumed to follow some distribution, $C_i \sim G$.

Along with lifetime, we observe $\delta_i = I(T_i \leq C_i)$, which gives us the necessary information to distinguish actual events from (right) censored observations, and $\boldsymbol{x}_i$, covariates on subject i. To summarize, we observe the triplet $(\widetilde{T}_i, \delta_i, \boldsymbol{x}_i)$ for each subject and wish to draw inference about T_i. For random censoring, we also observe a subject-specific entry time V_i, at which the subject is left truncated.

In order to do inference on T, the censoring must be *non-informative*, i.e. T_i

and C_i are assumed to be independent. This assumption typically means that if a subject is censored, we will not be able to say anything of its lifetime, except it being larger than the censoring time. If this is violated, estimations are severely biased and the direction of this bias depends on the situation.

2.5.3 Hazard and stratified hazard models

Since we are interested in the probability of survival, let $S_T(t) = 1 - F_T(t) = P(T > t)$ be the complement of the distribution function. Hence, this function has support in $\mathbb{R}^+$ with boundary condition $S(0) = 1$ and $S(\infty)$=0.

Next, consider

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \cdot \frac{1}{S(t)} = \\ = \frac{f(t)}{S(t)} \quad (1)$$

which is interpreted as the risk of having the event instantaneously after time t, given survival up to t. The quantity $\alpha(t)$ is called the *hazard* at time t. Note that $\alpha(t) \geq 0$ for $t \geq 0$, and can thus not to be intepreted as a probability.

By using the fact that $f(t) = \frac{\mathrm{d}F(t)}{\mathrm{d}t} = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}$, we get

$$\alpha(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t) \Longleftrightarrow S(t) = \exp\left(-\int_0^t \alpha(u)\mathrm{d}u\right) \quad (2)$$

From (1) and (2), important relations between hazard, density and survival are seen. These are useful for mathematical treatments of models, where it is necassary to move from one representation to another.

Formula (2) can also be expressed using the *cumulative hazard* as

$$S(t) = \exp\left(-A(t)\right),$$

where $A(t) = \int_0^t \alpha(u)\mathrm{d}u$. Note that this relation only holds true for continuous distributions.

Standard estimator of the survival function is the Kaplan-Meier estimator, defined for $t \in [0, \tau]$ and with assumption of non-informative censoring as

$$\widehat{S}(t) = \begin{cases} 1, & \text{if } t < t_1 \\ \prod_{t_i \leq t} [1 - \frac{d(t_i)}{Y(t_i)}], & \text{if } t \geq t_1, \end{cases}$$

where t_1 denotes the first event time, $d(t_i)$ is the number of events at event time t_i and $Y(t_i)$ is the number of subjects who potentially could have an event at t_i, i.e. $Y(t_i)$ is populated with all those subjects who have survived to a time just before t_i. ($Y(t_i)$ will later be defined as the risk set at t_i.) The quantity $d(t_i)/Y(t_i)$ is thus a conditional probability of having an event, given no event up to time t_i.

If the last observed time, τ, is a censoring time, the Kaplan-Meier estimator is not well defined for time points after this.

The variance of the Kaplan-Meier estimator is estimated by Greenwood's formula:

$$\widehat{V}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{t_i \leq t} \frac{d(t_i)}{Y(t_i)(Y(t_i) - d(t_i))}$$

Note that if $d(t_i) = 1$ for each event time t_i and all observations are complete, i.e. if events occur one at the time and there are no censoring, Greenwood's formula reduces to a standard binomial varance estimator.

The cumulative hazard, which provides a crude estimator of the hazard rate $\alpha(t)$ by studying its slope, is often estimated using the Nelson-Aalen estimator

$$\widehat{A}(t) = \begin{cases} 0, & \text{if } t < t_1 \\ \sum_{t_i \leq t} \frac{d(t_i)}{Y(t_i)}, & \text{if } t \geq t_1, \end{cases}$$

with variance estimated by

$$\widehat{V}[\widehat{A}(t)] = \sum_{t_i \leq t} \frac{d(t_i)}{Y(t_i)^2}$$

Here, the same assumptions as for the Kaplan-Meier estimator are used.

When fitting models to survival data there are essentially two families of models: accelerated failure time (AFT) models and proportional hazard models. For the former, (transformed) lifetime is modelled using a conventional linear model, e.g.

$$\log T_i = \boldsymbol{\beta}' \boldsymbol{x}_i + \varepsilon_i \Longleftrightarrow T_i = \exp(\boldsymbol{\beta}' \boldsymbol{x}_i) T_{0i},$$

where $\boldsymbol{\beta}'$ is the transposed vector of $\boldsymbol{\beta}$. Here a distribution for $\varepsilon_i = \log(T_{0i})$ is specified. Often used distributions for the error term, are log-normal, gamma and extreme value distributions (Allison [3]).

The other case focuses on modelling hazard and parametric assumptions are only imposed on for the effect of the predictors on the hazard. Cox [7] suggested a proportional hazard regression modell on the form

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) = \lambda_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_i x_{ik}), \qquad (3)$$

where $\boldsymbol{\beta}$ is a vector with regression coefficients, which we later want to estimate, and $\boldsymbol{x}_i$ is our covariate vector for individual i. $\lambda_0(t)$ is a baseline hazard and could be regarded as a benchmark hazard, in common for all individuals. (As mentioned above, this part of the model is non-parametric.) Here $\lambda_i(t) = Y_i(t)\alpha_i(t)$, where $\alpha_i(t)$ is definied in (1) and $Y_i(t)$ is a binary variabel indicating if subject i is at risk at t. We will refer to $\lambda_i(t)$ as the *intensity* at t.

The sum, $Y.(t) = \sum_{i=1}^{n} Y_i(t)$ is called the *risk set* and gives the number of subjects at a time point right before t, who potentially could have an event at t.

Note that $Y.(t_i)$ equals the set $Y(t_i)$, which was introduced earlier.

The baseline hazard, $\lambda_0(t)$, is not assumed to follow any particular distribution, but still has the properties of hazards, i.e. non-negative and real for $t \geq 0$. This implies that $\lambda_i(t)$ also have this properties.

By manipulating (3) we find that

$$\log\left(\frac{\lambda_i(t)}{\lambda_0(t)}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_i x_{ik},$$

where $\frac{\lambda_i(t)}{\lambda_0(t)}$ is the hazard ratio for i to the baseline.

A characteristic for the model (3) is that the hazard ratio for two subjects, i and i^*, is independent of time, since

$$\frac{\lambda_i(t)}{\lambda_i^*(t)} = \frac{\lambda_0(t)\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\lambda_0(t)\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i^*)} = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i^*)} = \exp\left(\sum_{i=1}^{k} \beta_i (x_i - x_i^*)\right)$$

hence proportional. This does not imply that the hazards for the subjects are constant over time, only that the ratio of their hazards is. The quantity is intepreted as the relative risk of having an event at t for a subject with covariates $\boldsymbol{x}_i$ compared to a subject with covariates $\boldsymbol{x}_i^*$.

We are also interested in letting the hazard differ across subgroups. For this reason, let

$$\lambda_l(t) = \lambda_{l0}(t)\exp(\boldsymbol{\beta}^T \boldsymbol{x})$$

be the hazard for strata $l = 1, \ldots, L$. Here, we let each strata have a unique baseline hazard but allow for risk parameter to have the same effect on covariates across different strata. By doing this, subjects in different subgroups are effected by covariates in the same way but the baseline may differ. (We will later relax this assumption in a competing risk setting.) This operation is necessary when there are evidence of non-proportionality for some covariate. One vital requirement is that the grouping of subjects is based on the past, not on future events.

When estimating parameters in the stratified Cox model, one maximizes the partial likelihood per stratum, as

$$L(\boldsymbol{\beta}) = \prod_l L_l(\boldsymbol{\beta}),$$

where

$$L_l(\boldsymbol{\beta}) = \prod_{j=1}^{n} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_j)}{\sum_{k \in R_{lj}} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_k)}$$

Here, $j = 1 \ldots n$ are ordered event times and R_{lj} is the risk set at event time t_j in stratum l. We have assumed here that no ties are present, an assumptions that need to be relaxed in our setting. Methods for handling ties are discussed in Hosmer and

Lemeshow [13]. We will use the method originally suggested by Efron [10], who let the partial likelihood be

$$L_l(\boldsymbol{\beta}) = \prod_{j=1}^{n} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{s}_j)}{\prod_{i=1}^{d_{lj}} \left(\sum_{k \in R_{lj}} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_k) - \frac{i-1}{d_{lj}} \sum_{k \in D_{lj}} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_k) \right)},$$

where $\boldsymbol{s}_j = \sum_{k \in D_{lj}} \boldsymbol{x}_k$ is a sum of covariate vectors and D_{lj} is the set of all subjects who experience the event from cause l at t_j and d_{lj} is the cardinality of this set. This method has shown to be more accurate than other methods, e.g. Breslow's method, and the difference is apparent when data contains a lot of ties.

In a extended version of (3), one lets the covariates be time-varying, thus violating the proportionality property. Furthermore predictions are diffucult to interpret since our covariates are, what Klein and Moeschberger [15] refer to as, *internal* as in subject specific. This means that, we would need to predict covariate paths (which might be possible in some cases and difficult or even impossible in other) for each customer in order to predict survival probabilities.

However, this inclusion might provide a more realistic model of customer behaviours, since information on customers do change over time. E.g. customers move, their premiums are adjusted etc. and this will effect the customer relationship.

2.5.4 Cumulative incidences

In many studies the probability of experiencing the event of interest is altered due to the presence of competing risks. In our case, we have defined five causes as to why a relationship potentially can end. Before proceeding with model development of hazard function, investigation on how these are effected by specific causes is needed.

This is done by considering the *cumulative incidence function*, given initial covariates $\boldsymbol{x}_0$,

$$P_{0l}(t|\boldsymbol{x}_0) = P(T \leq t, L = l|\boldsymbol{x}_0),$$

for definied competing risks, or causes of failure, $l = 1, \ldots, L$. Generally, survival data emerges from a two-state process, where an event either has happened or it has not. By splitting the state where it has happened, we arrive at the competing risk model, which is a Markov chain model. Each movement in this model is interpreted as the probability of having the event of interest before or at time t and that the observed failure is due to cause l. In the case of only one cause of failure, this probability is simply the distribution function. It is also possible to include more states into the Markov chain, one example being the three-state illness-death model. (See Crowder [8] for further details on competing risks models.)

We also define the overall survival probability as $P_{00}(t|\boldsymbol{x}_0) = P(T > t|\boldsymbol{x}_0)$, i.e. the probability of not having the event before time t, given initial covariates. (We use subindex 00 to distinguish the overall probability P_{00} from the cause specific event probabilities P_{0l}.) We will use the relation $P_{00}(t|\boldsymbol{x}_0) = 1 - \sum_{l=1}^{L} P_{0l}(t|\boldsymbol{x}_0)$ for this reason.

Next, define the cause-specific hazard as

$$\alpha_{0l}(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, L = l | T \ge t)}{\Delta t}, \quad l = 1, \dots, L$$

Letting $Y(t)$ denote the risk indicator, which allows for right censoring (and left truncation), we assume that

$$\lambda_{0l}(t) = Y(t)\alpha_{0l}(t), \quad \alpha_{0l}(t) = \alpha_{l0}(t)\exp(\boldsymbol{\beta}^T \boldsymbol{x}_l),$$

for $l = 1, \dots, L$. Here, $\alpha_{l0}(t)$ is again a unspecified baseline hazard for cause l and $\boldsymbol{x}_l$ is a cause-specific covariate vector, computed from $\boldsymbol{x}$ as described in 4.x. The second part of the right-hand side, $\exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)$, is the individual effect on the hazard and depends on the covariates of the subject. In this case we have assumed a multplicative form of this hazard, as in the Cox model in (3), but additive and mixtures of additive and multiplicative are also possible forms, see Scheike and Martinussen [19] for details on various hazard models.

Associated with this hazard is the integrated baseline hazard, which we encountered above in (2), $A_{l0}(t) = \int_0^t \alpha_{l0}(s)\mathrm{d}s$.

The parameter vector $\boldsymbol{\beta}$ is common for all causes, since the covariates are cause-specific, as mentioned in 2.5.3. If no effect is common for different stratas, this will give one fitted Cox model per stratum. The strategy of fitting different models for diferent stratas is, however, not sufficient when effects are alloweded to be identical across strata, as argued in Rosthøj, Andersen and Abildstrom [18]. The arguments are applied here, since the competing risks are similar in some senses.

Again, this (cause-specific) hazard is not a probability, but a rate and gives the risk of having an event by cause l in the next time interval, given no event (by any cause) up to a point in time.

Based on this quantity, we can express the cumulative incidence function as

$$P_{0l}(t|\boldsymbol{x}_0) = P(T \le t, L = l|\boldsymbol{x}_0) = \int_0^t \alpha_{0l}(s)S(s-)\mathrm{d}s,$$

where $S(s)$ was defined earlier as the survival function, with $S(s-) = \lim_{t \uparrow s} S(t)$. In terms of intensities, the survival function is expressed as

$$S(t) = \exp\left(-\int_0^t \alpha(s)\mathrm{d}s\right) = \exp\left(-\int_0^t \sum_{l=1}^{L} \alpha_{0l}(s)\mathrm{d}s\right),$$

where $\alpha(t) = \sum_{l=1}^{L} \alpha_{0l}(t)$ is the total intensity at time t.

Based on a sample $\{(\widetilde{T}_i, L_i, \boldsymbol{x}_i)\}_{i=1}^{n}$ of n observation - where $\widetilde{T}$ is observered failure time, only equal to actual survival time in case of event and L is failure variable ($L = 0$ if no event) - the cumulative incidence function is estimated based on the estimators $\hat{\boldsymbol{\beta}}$ and $\widehat{A}_{l0}(t, \hat{\boldsymbol{\beta}})$. These are in turn estimated from the stratified

Cox model and the underlying Nelson-Aalen estimator, respectively.

Order the event times (for any cause) as $0 = t_0 < t_1 < \cdots < t_m < \tau$, for events over the study interval $(0, \tau]$, and let $d_{lj} \geq 1$ denote the number of event of cause l occuring at t_j. The overall survival probibility is then estimated by

$$\widehat{P}_{00}(t_m|\boldsymbol{x}_0) = \prod_{j=1}^{m-1}(1 - \mathrm{d}\widehat{A}_0(t_j|\boldsymbol{x}_0)),$$

with $\widehat{A}_0(t_j|\boldsymbol{x}_0) = \sum_{l=1}^{L} \widehat{A}_{l0}(t_j, \hat{\beta}) \exp(\hat{\beta}^T \boldsymbol{x}_0^l)$. Here $\boldsymbol{x}_0^l$ is the cause specific covariate vector computed from $\boldsymbol{x}_0$. This estimator have the same conceptual interpretation as the Kaplan-Meier estimator, since survival up to a certain time is conditional on survival up to the time prior to this, and so forth. From this, an estimator of the cumulative incidence function can be derived as

$$\widehat{P}_{0l}(t_m|\boldsymbol{x}_0) = \sum_{j=1}^{m} \widehat{P}_{00}(t_{j-1}|\boldsymbol{x}_0)\mathrm{d}\widehat{A}_{l0}(t_j|\hat{\beta}) \exp(\hat{\beta}^T \boldsymbol{x}_0^l),$$

under the Cox model.

Technically, the competing risk model is as mentioned a special case of a multi state Markov model and the estimator $\widehat{P}_{0l}(t_m|\boldsymbol{x}_0)$ is derived from that setting in Anderson, Borgan, Gill and Keiding [4]. In this terminology each transistion out of the "active" state and into one of the "inactive" states (as in fig. 1) are absorbing. Thus, the transition probability matrix $\boldsymbol{P}(t)$, defined for $s < t \leq \tau$, is

$$\boldsymbol{P}(s,t) = \begin{pmatrix} P_{00}(s,t) & P_{01}(s,t) & \dots & P_{0L}(s,t) \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

where $P_{0l}(t)$, $l = 0, \dots L$ is the probability of moving from state 0 (the "active" state) into state l ("active" for $l = 0$ and "inactive" for $l = 1, \dots, L$) in the time interval $(s, t]$, conditional on being in state 0 at time s. All other probabilities are 0 or 1, because we then condition on being in an absorbing state. The corresponding transition intensity matrix, is given by

$$\boldsymbol{\alpha}(t) = \begin{pmatrix} -\sum_{l=1}^{L} \alpha_{0l}(t) & \alpha_{01}(t) & \dots & \alpha_{0L}(t) \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix},$$

since all transitions out of the initial state 0 are absorbing. Now, the Kolmogorov equation $\boldsymbol{P}'(t) = \boldsymbol{\alpha}\boldsymbol{P}(t)$ yields the differential equations

$$\frac{\partial}{\partial t}P_{00}(s,t) = -P_{00}(s,t)\sum_{l}\alpha_{0l}(t)$$
$$\frac{\partial}{\partial t}P_{0l}(s,t) = -P_{0l}(s,t)\alpha_{0l}(t),$$

with solutions:

$$P_{00}(s,t) = \exp\left(-\int_s^t \sum_{l=1}^{L}\alpha_{0l}(u)\mathrm{d}u\right),$$
$$P_{0l}(s,t) = \int_s^t \alpha_{0l}(u)P_{00}(s,u)\mathrm{d}u.$$

Again, the purpose here is to calculate $P_{00}(t) = P_{00}(0,t)$, the probability of not transiting into one of the absorbing states, by estimating the probabilities of these transitions over time.

3 Models and methods

As mentioned above, survival data is often subject to complex censoring schemes, suggesting other regression models than e.g. ordinary and logistic regression models. An appealing approach is the one developed by Aalen [1], who combined the theories of stochastic integration, martingales and counting processes into a methodology well suited for survival and event history data. For an introduction to the subject, see Therneau and Grambsch [21] and Anderson *et al.* [4].

This chapter will start with presenting survival analysis in a counting process setting, deriving the cumulative hazard function and survival function estimators, as well as the log-rank test.

3.1 Counting processes and survival analysis

Let $N(t), t \geq 0$ be a counting process, i.e. $N(t) \in \mathbb{Z}^+$ is a stochastic process with properties $N(0) = 0$, $P(N(t) < \infty, t \geq 0) = 1$ and with right-continuous paths and $\mathrm{d}N(t) \in \{0, 1\}$, where $\mathrm{d}N(t)$ denote the increment of the process over a (small) intervall $(t, t + \mathrm{d}t]$. This means that the process starts in 0 at some time of origin and has a jump of size 1 at a time point if there is one (and only one) event at this time point, and is unchanged if there is no event. Let the intensity, i.e. the rate of which the events in the process occur, be $\lambda(t)$, as defined in 2.5.3,

An example of a counting process is the (homogeneous) Poisson process, with intensity λ. Another would be, as in our case, customer churns in a insurance company where $N(t)$ counts the number of churns up to time t where churns occur with some intensity $\lambda(t)$.

For subjects $i = 1, \ldots, n$, we let the corresponding counting process be denoted by $N_i(t) = I(T_i \leq t, \delta_i = 1)$, yielding the aggregated counting process $N.(t) = \sum_{i=1}^{n} N_i(t) = \sum_{t_i \leq t} \delta_i$, for distinct event times t_i. $N_i(t)$ could be thougth of as a slow Poisson process, which start as 0 and at the subject specific event time has a jump of size 1.

Let the accumulated history up to time t of the customers be denoted by $\mathbb{F}_t$, which contains all information on customers gathered during the study period. This quantity is formally known as the (*self-existing*) *filtration* of the counting process and consists of σ-algebras, $\mathcal{F}_t$, generated by the (aggregated) counting process up to and including time t, i.e. $\mathbb{F}_t = \{\mathcal{F}_0, \ldots, \mathcal{F}_t\}$ where $\mathcal{F}_t = \sigma(N.(s), 0 \leq s \leq t)$. That is, all history or all information on the subjects, is contained in the filtration.

Under the assumption of independent censoring, we say that $N(t)$ is *adapted* to $\mathbb{F}_t$. Technically, this means that $N(t)$ is $\mathcal{F}_t$-measurable for each t. This assumption is automatically fulfilled in the type I and type II censoring schemes discussed in 2.5.2. If censoring is random the observed counting process is not adapted to the history and we need to enlarge the history with that generated by the censoring process, using a result known as the *innovation theorem*, see e.g. Aalen, Borgan and Gjessing [2] for discussion on this result. This follows from the fact that in type I and II censoring schemes, the censoring times are stopping times relative to the history $\mathbb{F}_t$ and the form of the intensity process is preserved even if we consider the larger history.

Besides these quantities, we need a definition of (continous) *martingales*. Let $M = \{M(t); t \in [0, \tau]\}$ be a stochastic process. M is a martingale relative to the history $\mathbb{F}_t$ if it is adopted to the history and it has the property

$$E(M(t)|\mathbb{F}_s) = M(s), \quad \text{for all } t > s.$$

These concepts are fundamental in stochastic calculus and probability theory, see e.g. Klebaner [14] for an introduction to the subject.

Given history up to a time right before t, i.e. given all information that is gathered and known up to this point, we have

$$E(\mathrm{d}N(t)|\mathbb{F}_{t^-}) = P(t \leq T \leq t + \mathrm{d}t|\mathbb{F}_{t^-}) = \lambda(t)\mathrm{d}t = Y(s)\alpha(s)\mathrm{d}t$$

Next, define the cumulative intensity $\Lambda(t)$ by $\int_0^t \lambda(s)\mathrm{d}s$ $[= \int_0^t Y(s)\alpha(s)\mathrm{d}s]$, for $t \geq 0$. This process has the property

$$E(N(t)|\mathbb{F}_{t^-}) = E(\Lambda(t)|\mathbb{F}_{t^-}) = \Lambda(t),$$

since $Y(t)$ is assumed to be predictable and thus fixed, given history up to t^-.

Let $M(t) = N(t) - \Lambda(t)$ be the *martingale counting process*, with increments 0, given history. This is seen from

$$\begin{aligned} E(\mathrm{d}M(t)|\mathbb{F}_{t^-}) &= E(\mathrm{d}N(t) - \mathrm{d}\Lambda(t)|\mathbb{F}_{t^-}) = \\ &= E(\mathrm{d}N(t)|\mathbb{F}_{t^-}) - E(\lambda(t)\mathrm{d}t|\mathbb{F}_{t^-}) = \\ &= 0. \end{aligned}$$

It can be shown that $M(t)$ is a martingale with respect to $\mathbb{F}$. The martingale property is satisfied, since for $s < t$

$$\begin{aligned} E(M(t)|\mathbb{F}_s) - M(s) &= E(M(t) - M(s))|\mathbb{F}_s) = \\ &= E\left(\int_s^t \mathrm{d}M(u)|\mathbb{F}_s\right) = \\ &= \int_s^t E(E(\mathrm{d}M(u)|\mathbb{F}_{u^-})|\mathbb{F}_s) = \\ &= 0. \end{aligned}$$

From the martingale counting process, we see the relation $N(t) = \Lambda(t) + M(t)$, with differential

$$\begin{aligned} \mathrm{d}N(t) = \mathrm{d}\Lambda(t) + \mathrm{d}M(t) =& \lambda(t)\mathrm{d}t + \mathrm{d}M(t) \\ =& Y(t)\alpha(t)\mathrm{d}t + \mathrm{d}M(t), \end{aligned}$$

where, as in 2.5.3, the intensity $\lambda(t)$ can be expressed as $\lambda(t) = Y(t)\alpha(t)$ with risk indicator $Y(t)$ and hazard $\alpha(t)$.

This is conceptually similiar to ordinary regression models; the left-hand side is the observed quantity, while the first term on the right-hand side is the systematic part of the model (which represents signal) and the second term is the random part (which represents noise). When $Y(t)$ is non-zero we have,

$$\frac{\mathrm{d}N(t)}{Y(t)} = \alpha(t)\mathrm{d}t + \frac{\mathrm{d}M(t)}{Y(t)}. \tag{4}$$

Conditional on $\mathbb{F}_{t^-}$, this gives, since $Y(t)$ is predictable,

$$\begin{aligned} E\left(\frac{\mathrm{d}M(t)}{Y(t)}|\mathbb{F}_{t^-}\right) &= \frac{E(\mathrm{d}M(t)|\mathbb{F}_{t^-})}{Y(t)} = 0, \\ \mathrm{Var}\left(\frac{\mathrm{d}M(t)}{Y(t)}|\mathbb{F}_{t^-}\right) &= \frac{\mathrm{Var}(\mathrm{d}M(t)|\mathbb{F}_{t^-})}{Y(t)^2} = \frac{E[(\mathrm{d}M^2(t))|\mathbb{F}_{t^-}]}{Y(t)^2} = \frac{\mathrm{d}\langle M\rangle(t)}{Y(t)^2} \end{aligned}$$

In the nominator in last quantity, $\langle M\rangle(t)$ is known as the *predictable variation process* (of $M(t)$), with increments

$$\mathrm{d}\langle M\rangle(t) = \mathrm{Var}(\mathrm{d}M(t)|\mathbb{F}_{t^-}) = \mathrm{Var}(\mathrm{d}N(t)|\mathbb{F}_{t^-}) = \mathrm{d}\Lambda(t)(1 - \mathrm{d}\Lambda(t)) \approx \mathrm{d}\Lambda(t),$$

which gives $\langle M\rangle(t) = \Lambda(t)$. We also mention the concept of *optional variation process* for a martingale, $[M]$. Formally, these two concepts are defined as:

$$\langle M\rangle(t) = \lim_{n\to\infty}\sum_{k=1}^{n}\mathrm{Var}(\Delta M_k|\mathbb{F}_{(k-1)t/n}),$$

and

$$[M](t) = \lim_{n\to\infty}\sum_{k=1}^{n}(\Delta M_k)^2,$$

where the time interval $[0,t]$ is partitioned into n subintervals of equal length t/n and $\Delta M_k = M(kt/n) - M((k-1)t/n)$. From the defintion of the optimal variation process, it follws that $[M] = N$.

Using these definitions along with that of martingales, we therefore have that

$$\mathrm{Var}(M(t)) = E(M(t)^2) = E\langle M\rangle(t) = E[M](t).$$

Now, letting $J(t) = I(Y(t) > 0)$, 1 if (at least one) subject is at risk at t and 0 otherwise, and using the convention 0/0=0, integrating (4) yields

$$\begin{aligned} \int_0^t \frac{J(s)}{Y(s)}\mathrm{d}N(s) &= \int_0^t J(s)\alpha(s)\mathrm{d}s + \int_0^t \frac{J(s)}{Y(s)}\mathrm{d}M(s) = \\ &= \int_0^t J(s)\mathrm{d}A(s) + \int_0^t \frac{J(s)}{Y(s)}\mathrm{d}M(s) \end{aligned} \tag{5}$$

which suggests the estimator

$$\widehat{A}(t) = \int_0^t \frac{J(s)}{Y(s)} \mathrm{d}N(s)$$

for the (stochastic) integrated hazard function, $\int_0^t \alpha(s)\mathrm{d}s$, and is refered to as the Nelson-Aalen estimator.

Ideally, we seek the estimator of $A(t)$, which is not possible since there is a (depending on sample size) probability that $Y(s) = 0$. But if this probability is small then $J(s) = 1$ for most s and $A^*(t) = \int_0^t J(s)\alpha(s)\mathrm{d}s \approx \int_0^t \alpha(s)\mathrm{d}s$. Using these two notations, reformulate (5) as

$$\widehat{A}(t) - A^*(t) = \int_0^t \frac{J(s)}{Y(s)} \mathrm{d}M(s) \tag{6}$$

The heuristic behind our derived estimator is now: because the process $M(t)$ is a martingale (with zero mean), the stochastic integral $\int_0^t \frac{J(s)}{Y(s)}\mathrm{d}M(s)$ is also a martingale when the integrand is predictable. Thus the estimator is unbaised, i.e. $E[\widehat{A}(t) - A^*(t)] = 0$, and it can be shown by applying the martingale central limit theorem that, under regularity conditions, $n^{1/2}(\widehat{A}(t) - A^*(t))$ converges in distribution towards a Gaussian martingale for $t \in [0, \tau)$[7]. An unbiased estimator of the variance, σ^2, is then obtained by applying the optimal variation process:

$$\hat{\sigma}^2(t) = \int_0^t \frac{J(s)}{Y^2(s)} \mathrm{d}N(s) \tag{7}$$

Practically, the Nelson-Aalen estimator is a sum calculated over the distinct event times of the process, $\widehat{A}(t) = \sum_{t_i \le t} \frac{1}{Y(t_i)}$, since integrating over a counting process is equivalent to summing the integrand over the jump times. The estimator has already been stated in 2.5.3. Note that here $d_i \equiv 1$, i.e. the event times are distinct, whereas in 2.5.3 $d_i \ge 1$, and ties could be dealt by using e.g. Efron's approximation.

The derivation of a survival function estimator appeared in 2.5.3 and is restated here: Kaplan-Meier estimator is given by

$$\widehat{S}(t) = \prod_{t_i \le t} (1 - \frac{1}{Y(t_i)}) = \prod_{t_i \le t} (1 - \Delta\widehat{A}(t_i)),$$

where $\Delta\widehat{A}(t) = \widehat{A}(t) - \widehat{A}(t-)$ is the increment in the Nelson-Aalen estimator. Note that this estimator is only valid for (absolutely) continuous distributions. For general distributions, one needs to apply product-integrals to estimate the survival function. Furthermore, one can show that $n^{1/2}(\widehat{S}(t) - S(t))$ asymptotically is dis-

[7]See [4] for formal derivation and formulation of the martingale CLT.

tributed as a men zero Gaussian process, with variance estimated by

$$\widehat{\tau}^2(t) = \widehat{S}(t)^2 \widehat{\sigma}^2(t) = \widehat{S}(t)^2 \sum_{t_i \le t} \frac{1}{Y(t_i)^2}$$

with $\widehat{\sigma}^2(t)$ as in (7). Alternatively, this is replaced by

$$\widetilde{\tau}^2(t) = \widehat{S}(t)^2 \sum_{t_i \le t} \frac{1}{Y(t_i)(Y(t_i) - 1))}$$

From this we can create (point-wise) confidence interval for the Kaplan-Meier estimator; a standad $100(1-\alpha)\%$ confidence interval for $\widehat{S}(t)$ has the form $\widehat{S}(t) \pm z_{1-\alpha}\widehat{\tau}(t)$.

For testing equality of hazard functions between subgroups of a population, a standard approach is the log-rank test. One is then interested in testing the hypothesis

$$\mathrm{H}_0 : \alpha_1(t) = \alpha_2(t) = \cdots = \alpha_k(t) \quad \text{for } t \in [0, \tau],$$

for k subgroups. Let $N. = \sum_{h=1}^{k} N_h$ and $Y. = \sum_{h=1}^{k} Y_h$ be the aggregated counting process and risk set, respectively, of all subgroups. Introduce a nonnegative and predictable weight process $K(t)$, which is 0 whenever $Y.$ is 0. Define the process

$$Z_h(t) = \int_0^t K(s)\mathrm{d}N_h(s) - \int_0^t K(s)\frac{Y_h(s)}{Y.(s)}\mathrm{d}N.(s),$$

for $h = 1, \ldots k$, which can be showned to be zero mean martingales under the null hypothesis. An unbiased estimator of the covariance for two of these processes is given by

$$V_{hj}(t) = \int_0^t K^2(s)\frac{Y_h(s)}{Y.(s)}\left(\delta_{hj} - \frac{Y_j(s)}{Y.(s)}\right)\mathrm{d}N.(s),$$

where δ_{hj} denotes the Kronecker delta, equal to 1 when $h = j$, 0 otherwise. The Z_h's can be seen as observed-minus-expected processes, weighted by some process K. (The expected term is not really an expected number in its true sense, since that term is stochastic, but under the null its) See Aalen, Borgan and Gjessing [2] for more information on nonparametric testing.

Since the sum of these processes, $\sum_h Z_h(t)$, is 0 we introduce a vector of dimension $(k-1)$, $\boldsymbol{Z}(t) = (Z_1(t), \ldots, Z_{k-1}(t))^T$ to test the hypothesis. In addition we need the covariance matrix $\boldsymbol{V}(t) = \{V_{hj}(t)\}_{h,j=1}^{k-1}$. With these two matrices we create the test statistic

$$X^2(t) = \boldsymbol{Z}(t)^T \boldsymbol{V}(t)^{-1} \boldsymbol{Z}(t),$$

which, under the null, approximately follows a chi-squared distribution (with $k-1$ degrees of freedom).

Different choices of $K(t)$ results in different tests, and the log-rank test is equivalent to setting $K(t) = I(Y.(t) > 0)$.

Since it is often of interest to control for other potentially important factors when comparing differences of survival times across the primary covariate, one extend the results above to stratified analysis. The null stated above would then be reconstructed as

$$\mathrm{H}_0 : \alpha_{1l}(t) = \alpha_{2l}(t) = \cdots = \alpha_{kl}(t) \quad \text{for } t \in [0, \tau] \quad l = 1, \ldots, L$$

When creating a test statistic we add an index to the Z_h's and V_{hj}'s to indicate the dependence on the stratum. That is, we let

$$\begin{aligned} Z_{hl}(t) &= \int_0^t K_l(s)\mathrm{d}N_{hl}(s) - \int_0^t K_l(s)\frac{Y_{hl}(s)}{Y_{\cdot l}(s)}\mathrm{d}N_{\cdot l}(s), \\ V_{hjl}(t) &= \int_0^t K_l^2(s)\frac{Y_{hl}(s)}{Y_{\cdot l}(s)}\left(\delta_{hj} - \frac{Y_{jl}(s)}{Y_{\cdot l}(s)}\right)\mathrm{d}N_{\cdot l}(s), \end{aligned}$$

This gives a test statistic that aggregate information over L strata:

$$X^2(t) = \left(\sum_l \boldsymbol{Z}_l(t)\right)^T \left(\sum_l \boldsymbol{V}_l(t)\right)^{-1} \left(\sum_l \boldsymbol{Z}_l(t)\right),$$

which, again, is approximately chi-squared distributed with $k-1$ df under the null.

The log-rank test is a powerful test against an alternative of proportional hazard, or the Cox model assumption, and detects consistent differences in survival times over the study period.

3.2 Cox model

For regression models in survival analysis, assume we have counting processes $N(t) = (N_1(t), \ldots, N_n(t))^T$ with intensities $\lambda(t) = (\lambda_1(t), \ldots, \lambda_n(t))^T$, as in (3). The intensities under a Cox model are expressed as $\lambda_i(t) = Y_i(t)\alpha_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i(t))$, with covariates $\boldsymbol{x}_i(t)$ and where $Y_i(t)$ is a risk indicator - 1 if subject i is at risk at time t, 0 otherwise. As mentioned in 2.5.3, we use the partial likelihood the find estimate of $\boldsymbol{\beta}$, by, for each distinct event time, comparing the covariates of the subject having an event with those who are in the risk set just prior to the that event time and multipying over all observed event times. That is, we find values of β_i to maximize

$$L(\boldsymbol{\beta}) = \prod_{t_j} \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i(t_j))}{\sum_{i=1}^n Y_i(t_j)\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i(t_j))} = \prod_{t_j} \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i(t_j))}{\sum_{R_j}\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i(t_j))},$$

for event times t_j. Here R_j is the risk set at t_j. Let

$$S^{(0)}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t)) = \sum_{R_t} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t)),$$
$$\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_i(t) \boldsymbol{x}_i(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t)) = \sum_{R_t} \boldsymbol{x}_i(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t)),$$
$$\boldsymbol{S}^{(2)}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_i(t) \boldsymbol{x}_i(t)^{\otimes 2} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t)) = \sum_{R_t}^{n} \boldsymbol{x}_i(t)^{\otimes 2} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t)),$$

where $\boldsymbol{x}^{\otimes 2}$ is the Kronecker product of the column vector $\boldsymbol{x}$. The logarithm of the partial likelihood can then be expressed as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \{\boldsymbol{\beta}^T \boldsymbol{x}_i(s) - \log S^{(0)}(\boldsymbol{\beta}, s)\} \mathrm{d}N_i(s)$$

The score function then becomes

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{x}_i(s) - \frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, s)}{S^{(0)}(\boldsymbol{\beta}, s)} \right\} \mathrm{d}N_i(s)$$

Solving $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = 0$, gives the estimate $\widehat{\boldsymbol{\beta}}$. See Scheike *et al.* [19] for different approaches for deriving this result.

For large samples, it can be shown that, under certain conditions, the maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ is approximately multivariate normal around the true $\boldsymbol{\beta}$ and with covariance matrix estimated by inverse of observed information matrix. Under the Cox model, the observed information matrix is equal to

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\beta}) &= \left\{ -\frac{\partial^2}{\partial \beta_h \partial \beta_j} \log L(\boldsymbol{\beta}) \right\} = -\frac{\partial}{\partial \boldsymbol{\beta}^T} \boldsymbol{U}(\boldsymbol{\beta}) \\ &= \int_0^{\tau} \left\{ \frac{\boldsymbol{S}^{(2)}(\boldsymbol{\beta}, s)}{S^{(0)}(\boldsymbol{\beta}, s)} - \left(\frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, s)}{S^{(0)}(\boldsymbol{\beta}, s)} \right)^{\otimes 2} \right\} \mathrm{d}N.(s), \end{aligned}$$

and the covariance matrix is estimated by $\boldsymbol{I}(\widehat{\boldsymbol{\beta}})^{-1}$.

Using these quantities, we eaily arrive at some familiar test statistics for hypothesis testing. Especially we have for tests on the form $\mathrm{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$

$$\textit{Likelihood ratio test: } \chi^2_{LR} = 2\{\log L(\widehat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta}_0)\}$$

$$\textit{Score test: } \chi^2_{SC} = \boldsymbol{U}(\boldsymbol{\beta}_0)^T \boldsymbol{I}(\boldsymbol{\beta}_0)^{-1} \boldsymbol{U}(\boldsymbol{\beta}_0)$$

$$\textit{Wald test: } \chi^2_{W} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \boldsymbol{I}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

All of these test statistics are asymptotically equal and under the null hypothesis, chi-squared with p (the number of parameters in the model) degrees of freedom. Then we estimate the cumulative baseline hazard, $A_0(t) = \int_0^t \alpha_0(s)\mathrm{d}s$, by first noting that the intensity of the aggregated counting process $N.(t) = \sum_i N(t)$ is on the form

$$\lambda.(t) = \sum_i \lambda_i(t) = \alpha_0(t)\left(\sum_i Y_i(t)\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t))\right)$$

With known $\boldsymbol{\beta}$, this would be a multiplicative intensity process, in which case $A_0(t)$ would be estimated by

$$\widehat{A}_0(t, \boldsymbol{\beta}) = \int_0^t \frac{\mathrm{d}N.(s)}{\sum_i Y_i(s)\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(s))}$$

By replacing $\boldsymbol{\beta}$ with its estimate, we arrive at the *Brelsow estimator* of the cumulative baseline hazard:

$$\widehat{A}_0(t) = \int_0^t \frac{\mathrm{d}N.(s)}{\sum_i Y_i(s)\exp(\widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_i(s))} = \sum_{t_j \leq t} \frac{1}{\sum_{R_j}\exp(\widehat{\boldsymbol{\beta}}^T \boldsymbol{x_i}(t_j))}$$

Given fixed covariate values, $\boldsymbol{x}_t(0)$, at time of origin, we have the estimate $\widehat{A}(t|\boldsymbol{x}_0) = \exp(\widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_0)\widehat{A}_0(t)$. In a stratified version of the Cox model, the results here extend as indicated in 2.5.3.

An extension of the Cox model is a generalization of the proportional hazard model (3) by letting covariates, some or all, as well as effects vary over time. This will, of course, make the assumption of proportional hazard invalid.

In a Cox model with time-varying covariates and time-dependent effects, the (mulitplicative) intensity for subject i is modelled as

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp(\boldsymbol{\beta}(\boldsymbol{t})^T \boldsymbol{x}_i(t)), \tag{8}$$

where $\boldsymbol{x}_i(t)$ is a p-dimensional bounded covariate vector. Both $Y_i(t)$ and $\boldsymbol{x}_i(t)$ are predictable, which means that both are fixed just prior to t. We will however only consider the case of fixed covariates, i.e. $\boldsymbol{x}_i(0)$.

The model in (8) represents an extremely flexible type of models, and a semi-parametric version of it, suggested by Martinussen, Scheike and Skovgaard [17], is on the form

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp(\boldsymbol{\beta}(t)^T \boldsymbol{x}_i(t) + \gamma^T \boldsymbol{z}_i(t)) \tag{9}$$

Here, $\boldsymbol{z}_i(t)$ is a q-dimensional covariate vector with time invariant effect vector γ, while $\boldsymbol{x}_i(t)$ and $\beta(t)$ now has dimension $p - q$. By letting $\beta(t) \equiv \beta$ and $\boldsymbol{x}_i(t) \equiv \boldsymbol{x}_i$, we have the model stated in (3).

The idea is to use the model in 9 for testing the Cox model assumption of proportional hazard, specifically if its effects are time invariant. If this is true for at least one effect, the Cox model is misspecified and need some more care.

Inference in this model, as well as other dynamic hazard models, is based on the cumulative regression function

$$\boldsymbol{B}(t) = \int_0^t \boldsymbol{\beta}(s)\mathrm{d}s,$$

where the individual regression coefficients β_j are "smooth" enough to be integrated. If we suspect time varying effects, we are interested in finding deviations from the (time invariant) Cox model. For that reason, express the regression coefficients as

$$\beta_j(t) = \beta_j + \theta_j g_j(t),$$

where $g_j(t)$ is some know function. One is then interested in testing the hypothesis $H_0 : \theta_j = 0$, for all j. A standard choice is $g_j(t) = \log(t)$ and by using the score function and the observed information matrix one can evaluate a test statistic for the devation from the Cox model, see Martinussen and Sckeike [19] for details.

Formal tests for testing time-varying effects includes the Kolmogorov-Smirnov test (KS) and the Cramer-von Mises (CvM) test. The KS test is a supremum kind of thest between the differences of grouped counting processes, weighted by appropriate quantities, such as the group-specific risk sets. The alternative CvM test is a infinium kind of test of the differences between the grouped counting processes. Both of the them test whether or not one can reject the null hypothesis of linear trend $\boldsymbol{B}(t) = \boldsymbol{\beta}t$.

3.3 Method

Using the concepts from 2.5, we study the cumulative incidence functions on product and customer level for focus customers. Incidences are the result of some cause and we want to estimate survival time by first estimating the hazard for each of the causes given covariates.

Survival time is here the number of churn-free time units (months) from time of origin to end of study after 35 time units. The initial risk set, $Y.(t_0)$, consists of those customers with at least one specific focus product at t_0. The risk set at the end of the study, $Y.(\tau)$, consists of those that have not had an event so far. Customers are thus subject to right censoring and left truncation.

As described in 2.1, there are two types of models being considered: time-invariant and dynamic. The first is based on information given at the time of origin. The second uses the same information while allowing for effects to be time-varying.

The analysis is carried out using the statistical softwares SAS and R. Especially **proc phreg** - which is a flexible procedure for estimating proportional hazard regression models and testing e.g. hypothesis on the regressors - in SAS and the **survival** - which contains a number of non-parametric estimates and tests, for example the Kaplan-Meier estimate and the log-rank test - and **timereg** - which allows for estimation of dynamic regression models of the sort in (8) as well as additive models - libraries in R are used. Also, additional SAS macros are used to estimate cumulative incidence functions and to score customers, see Rosthøj *et al.* [18]. Data manipulation is done is SAS.

The predictive models are produced in three step:

1. Based on a stepwise selection method with significance level of confidence limits set to near 1, we produce a sequence of models - ranging from the null to the full model - by minimizing the AIC statistic at each step.

2. From this sequence of models, we compare in detail those in a neighbourhood of the optimal model (the on with lowest AIC) by switching to a best subset selection approach. The neighbourhood is of arbitrary size and depends on whether or not there exists a unique optimal model. It is created by letting the number of included covariates vary over an interval including the number of the optimal model.

3. From the list of best subsets models, we choose which ever one - not necessary the originally optimal - has the lowest AIC.

This method is suggested in Shtatland, Kleinman and Cain [20]. The AIC is defined as

$$AIC = -2 \log L_C + 2p,$$

where p is the number of covariates introduced into the model and L_C is the partial likelihood produced in the Cox model. By including p covariates in the model, the sequence will produce Lp models (where L is the number of causes defined), which is a manageble number compared to the possible 2^{Lp} models.

In step 2. we consider models with number of included covariates near the optimal model, i.e. over an interval $[Lp - c_1, Lp + c_2]$ for some constants c_1, c_2. The reason for choosing a best subset selection approach in this step instead of stepwise or forward/backward selection is due to the relative unstable estimates that are produced in those methods. Also, since the number of possible covariates grow rapidly because of the competing risk setup, we need methods to manage this.

In addition, we will test for equality of cause-specific covariates across risk strata. For insignificant differences we will refit the model with these redefined combined covariates, described in 3.2.

In our study, we have introduced a lot of ties. While we proposed Efron's method for dealing with these in section 2.5.3, not all software features this method and assumes event to happen one at the time. Thus, we need to break these ties and this is done by adding random (uniform) noise to end points of the intervals where an event occurs.

Since time in this study refers to calender time, one need to be observant about this when modelling. Due to the construction of policies, there is a natural end date of a version of a policy, at which the cancellation take effect. This typically means that churns will occur (in most cases) in the same calender month as it was originally first signed, i.e. churns occur during renewal period. This impose a certain structure on the model, which is handled by stratify on renewal month.

In figure 2, the survival function is estimated for two subpopulations among motor focus customers: those with renewal in January and June, respectively. Studying the graphs, there seems to be a cyclic trend where there are sudden drops in the curves. For customers with renewal in January, there are drops in the curve after 12 and 24 months and similarly 6, 18 and (in smaller extent) 30 for customers with renewal in June. The shapes of the curves are the same, however, which suggests that there is no difference in the risk itself between subpopulations, but that there is a shift in time. This effect that arises due to our study design is as mentioned, solved by letting each subpopulation has its own baseline hazard but letting them share individual effects, as described in 2.5.3.

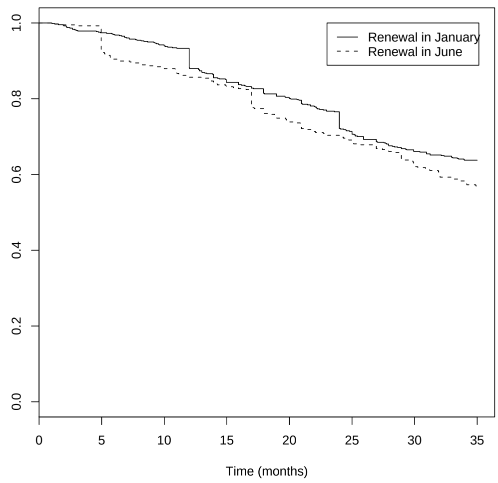


Figure 2: *Kaplain-Meier estimates stratified on expiration months for motor policy holders for January and June.*

The same effect can also been seen from table 2 below, where we compare the fraction of churns (C) within a calender month (1, ..., 12) to those with renewal (R) in that month. (Note that each row sum to 100%, since we consider all churns in observed in that month.)

Especially, we see the that for all motor policy holder that churned in the month of January (either in 2005, 2006 or 2007), nearly 50% of them had renewal in the same month, but only 4% had renewal in the next month.

Our predictive model will involve one more stratification, in order to handle the competing risks. This is done by stacking stacking covariates, with as many copies or stacks of the original data set (with n observations) as the number of causes used. Each stack has a numeric stratum indicator, $l = 1, \ldots, L$, created by letting the first n observations corresponding to the first cause of cancellation have $l = 1$ and the next n correspondning to the second cause have $l = 2$, and so forth. For

	R 1	2	...	11	12
C 1	47%	4%	...	5%	6%
2	11%	41%	...	5%	5%
⋮	⋮	⋮	⋮	⋮	⋮
11	8%	5%	...	28%	5%
12	6%	4%	...	8%	32%

Table 2: *Percentage of cancellations (C) in renewal months (R) among motor customers during 2005-2007*

each covariate effective for a specific cause, a cause-specific covariate is included in the stacked data. If Z is effective for cause 1, then we include the new covariate $Z_1 = Z$ while $l = 1$ (i.e. for the first n observations) and $Z_1 = 0$ for $l \neq 1$. If Z is effective for cause 2 as well, we let $Z_2 = Z$ while $l = 2$ (i.e. for observation $n+1$ to $2n$) and $Z_2 = 0$ for $l \neq 2$.

The idea here is to use the concept of non-informative censoring when letting each copy of the data correspond to one cause. When a event occur due to a cause other than that of interest, we set it as right censored and assume that this would not imply any sort of knowledge on potential events after, would this first, "uninteresting", event not had occured.

If the effect is equal in both stratas, let $Z_{12} = Z$ for $l = 1, 2$ and 0 otherwise. If no effects are equal over stratas, these stratified Cox models are equivalent to fitting L seperate Cox models. But applying the latter fitting strategy would not allow us to test equality of covariates across strata.

4 Data

The data is made out of two parts: historical and present data. The historical part consists of data on claims frequency and exposure prior to study. The present data contains specific information such as age, number of insurances of each product, at the beginning of the study. Note that, with respect to our definitions, most of these data is on household level. However, some data is on primary customers, e.g. age. Usually, only one member of a household is subject to campaigns and thus the primary customer has to fulfill communiction and hygiene rules set for specific campaigns.

A complete overview of used data is seen below.

Index	Covariate	Description
1	ANC	Survival time, calender time (in months)
2	D	Censoring variable, 1 if event, 0 otherwise
3	TYPE	Churning cause, I.-V. if D=1, 0. otherwise
4	EXPMONTH	Expiration month on primary policy
5	ANCPRE	Duration (in months) prior to study
6	AGE	Age of primary customer at beginning of study
7	GROUP	1 if customer has membership in some organisation
8	FORSAAR	Total exposure over preceding 4 year period
9	CLAIM	Number of claims reported over preceding 4 year period
10	CLAIMPCT	Claim ratio over preceding 4 years
11	FAM	1 if at least 1 home insurance, 0 otherwise
12	HUS	1 if at least 1 house insurance, 0 otherwise
13	OLY	1 if at least 1 accident insurance, 0 otherwise

Table 3: *Data on customers given at beginning of study.*

Covariate 1-3 are the observed results during the study. Covariate 4 is used as stratification variable and divide the population into 12 subgroups, depending on renewal months.

Covariate 5, ANCPRE, refers to how long a customer has had a specific product prior to the study. When modelling e.g. duration among motor policy holders, this covariate states how long a customers has been a motor policy holder.

Covariate 7 could be interesting, since membership in organisation with some contract with the insurance company can entitle discounts or other benefits or special service. Some customers have more than one membership, so without fully investigating how different memberships effect duration we simply group on none or at least one membership.

Covariate 8 measure the total exposure of the customer during the last 4 years, i.e. the total amount of time units which the insurance company has been exposed to some risk. A policy (no matter what line of business) contributes with 1, if it runs for 1 year. For each customer, the total exposure is thus (strictly) positive, with a highest observed value of 51, which is equivalent to about $(51/4 \approx)11$ policies on average per year.

Covariate 9 ranges from 0 to 20, and gives the number of reported claims during the last 4 years. This variable is regrouped, by grouping customer with 1-3 claims and customer with more than 3 claims into CLAIM1 and CLAIM2, respectively. CLAIM0 is the group of customers without any reported claims. Note that these are not product-specific claim and could be of any sort and size.

Covariate 10 is calculated as the contribution to the RBNS reserve of claims reported during the last 4 years (the same historical time period as covariate 8 and 9) over net earned premiums. This claims-to-premiums ratio is set to 0 if there are no claims reported or if the claim size or net earned premium is negative. The latter case occur if there were claims prior to this 4-year period, the of size which were overestimated and then recalculated in the 4-year period. The premiums could also be negative, in similar fashion, if the policies were cancelled before renewal and the premiums were accounted for in the previous period. Both of these situations occur due to accounting principles and by setting the ratio to 0 in these cases, although not entirely accurate, disregards claims dated prior to the 4-year period. As with covariate 9, CLAIM, this is not product-specific, but rather an overall ration for specific customers. It ranges from 0 to 60, for a few extreme observations. Everyone over 1 are all the customers that the insurer have had more expenses on than income from.[8]

Covariates 11-13 tell us about the focus product portfolio of the household at the beggining of the study. FAM and HUS can initially be treated as two seperate products, as customers can purchase both of them, either if the live in a house (in which case they also might need a house insurance) or a flat.

As mentioned earlier, our data is subject to type I censoring, see 2.5.2. The first monthly data is measured at the date 31.01.2005, time of origin t_0 , and the last at the date 31.12.2007, called τ, giving a study period of 36 months. Events after τ are considered censored. For a customer to be in the risk set at t_0, there has to be at least one active product-specific policy in the portfolio at that time point.

By defining the risk set at the end of each month, we will make the risk set for the (beginning of the) next month predictable. Practically this means that in a time interval, e.g. a month, the risk set is created first and then churns are excluded from this set at the end of the interval. This will create semi open time intervals, with left end points open and right points closed, in line with the counting process format described below.

In addition to the monthly data, which has a repeated measurement setup, we have an overview table, which is used for controlling the observations in the primary data. This overview table is on product level and is created with respect to our definition of focus products and customer relationship characteristics, see definitions in 2.3. This means that for each customer we have the latest coherent relationship, which might have been preceded by a relationship that was disrupted for more than

[8]Claims ratio, or claims percentage, is a important ratio on an aggregated level for insurance companies when measuring the profitability of the product portfolio. In particular, the claims ratio is one part when calculating the COR (combinded operating ratio), the second being that of expenses over written business.

12 months, thus considered churned. Also, the table gives information on when the policies originally were signed and - if they were - cancelled.

This control is important when looking at the monthly data, in which a customer can have zero focus products, but are not removed when the customer return with at least one focus product within 12 months. This allows temporarily interruptions in line with the company's definitions.

We then divide this table into tables corresponding to each focus product and conclude which customer has which focus product during our study period and how many. By letting the first signing date define the start and the last cancellation date define the end of the corresponding relationship, per focus product, we have created coherent intervals for each customer wherein the customer is a focus customer. If this interval overlaps the study period, we keep the customer in the study.

5 Results

Here results from the analysis on motor customers are presented. We start by presenting some of the univariate analysis, for determining which covariates might influence survival time, if we disregard the competing risks for the moment. In 5.2 Model selection we present a multivariate Cox model to use as reference when calculating cumulative incidence functions in 5.3.

5.1 Covariate analysis

From a total of roughly 85 000 customers (those in table 1, p. 8) who had at least one active motor policy at the beginning of the study period, a sample of 5 000 where used for the analysis. We used a simple random sample, while controlling for the stratification variable EXPMONTH.

Due to the large size of the sample, the same set were used for the univariate analysis.

By comparing the distribution of age among churned customers to that of those who didn't, we come up with a reasonable grouping of age. The Kaplan-Meier estimate for this grouping is seen in figure 3(a). This suggests that younger customer have a higher churning probability than the other groups. Age group 55-70 has the highest probability, albeit 70+ is not very far off. (There are some reasons as to why we want to keep these groups seperated through the analysis, and they will thus be threated as two distinct groups.) It is seen that among 20-39 years old customers, ca 40 % is still in the risk set after 3 years, compared to ca 75 % among 55-70 years old. This is supported when performing a log-rank test, which gives a test statistic $X^2(\tau) = 131$ on 3 degrees of freedom under the null of no difference in survival bewteen age groups - a highly significant result. Extending the test to its statrified version, yielded a test statistic on 164 on 3 df.

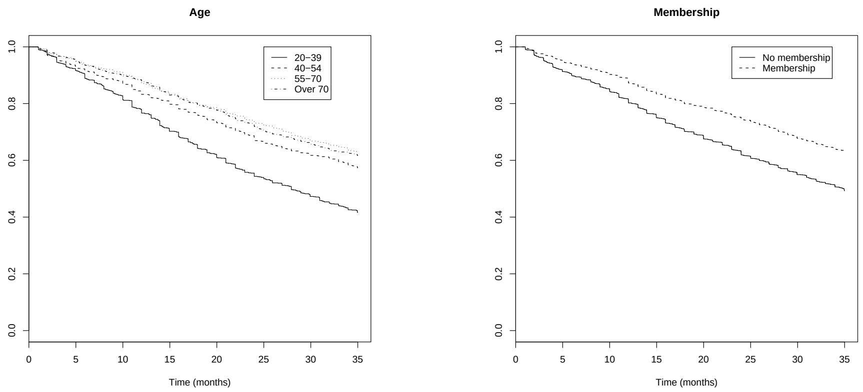
When examining the effect of membership, we introduced a dummy variable, 1 if the household is a member in at least one partner, 0 otherwise. This is a rather unprecise classification, but is a reasonable start.

Performing a log-rank test for the two groups, we found a test statistic on 99.2 on 1 df, corresponding to a p-value near 0 and we conclude that there is a difference in survival times between them.

A stratifed version of the test, controlling for the renewal month variable, EXPMONTH, gives the same conclusion, based on a slighlty higher chi-squared test statsitc.

Using a Kaplan-Meier estimate, it is obvious that customers with membership are more likely to renew their motor policies. Ca 50% of non-members have churned during the 3 years, while only ca 25% of members, as seen in figure 3(b).

When examining these two variables, one finds that among members, ca 40% belong to age group 55-70 and another 33% to age group 70+. In other words, it is more common among older customers to have a membership, so there seems to be an interaction here.



(a) (b)

Figure 3: (a) *Kaplan-Meier estimates for age groups among motor policy holders.* (b) *Kaplan-Meier estimates for membership among motor policy holders.*

To examine the effect of prior product-specfic duration to survival, we use a penalised smooting spline with 4 degrees of freedom in a Cox model with this as the only covariate. That is, we perform estimation of the form

$$\widehat{\alpha}(t|x) = s(x),$$

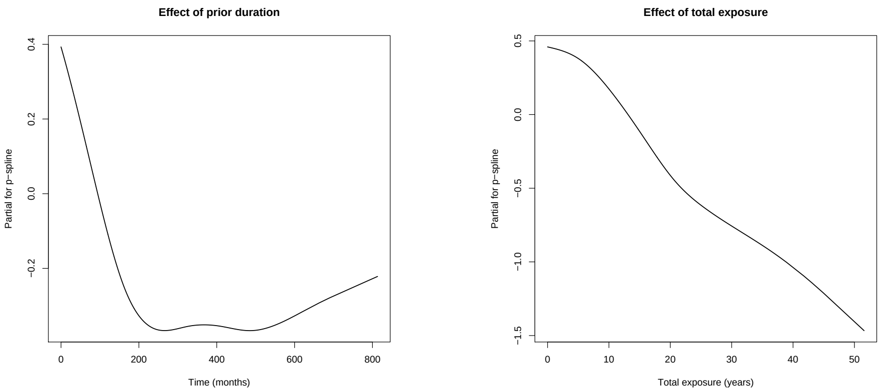
where x is the the covariate at hand and $s(x)$ is a penalised smoothing spline that takes the covariate as argument. (See e.g. Hastie, Tibshirani and Friedman [12] for details on non-parametric smoothing splines.)

When plotting this predictor against the outcome, as seen in figure 4, we get a feeling for the functional form of the covariate.

The same procedure is used for the covariate FORSAAR, which is the total exposure of the household in a 4-year period prior to the study. With the influence on the y-axis, it seems to have a decreasing effect on survival as the prior duration increases, up to a point at abut 20 years (=240 months) after which the effect flats out. This would imply that the risk of churning is decreasing the longer the customer relationship is active, at least up to a certain point. After this point, there is not enough data to support the increasing trend. It is however plausible to think that after such along time, the customer has reached a certain age and might no longer need a car.

From the other graph, one can conclude that the risk decreases lineary as the exposure increases, i.e. customers with more policies a more likely to stay. Again, data is skewly distributed over the range of exposure, with highest concentration in between 10 and 20 exposure years, which corresponds to ca 2-5 policies on average per year.

While disregarding the tails in the graphs in figure 4, where data is sparse, the seem to be a log-linear trend in ANCPRE and FORSAAR, which would suggest that we could model model them directly in a Cox model. We will however return to this in the next section.



(a) (b)

Figure 4: (a) *Effect of prior duration.* (b) *Effect of total exposure*

Log-rank test of covariate CLAIM, gives a borderline significant effect of the number of claims reported prior to the study on survival time. If assuming no difference between the different classes, we over estimate the survival of those customers without any claims.

Class	n	Observed	Expected
0 claims	1870	846	804
1: 1-3 claims	2405	997	1053
2: $\geq$4 claims	705	319	304
	Chisq=5.9	on 2 df,	p=0.05

Table 4: *Stratified log-rank test for claim groups, when controlling for renewal month*

By plotting the survival estimator for the claim groups, see 5, the same result is seen. The other plots in the figure, shows the effect on survival time of having other focus product than motor. E.g. those customers that have home insurance (FAM), in addition to a motor insurance, at the beginning of the study seem to be more likely to stay with their motor policy. The same thing for house insurance (HUS), and also for accident (OLY) to a smaller extent.

When modelling the effect of other insurance (FAM, HUS and OLY), there seem to be some deviations from Cox model assumption, when testing for violation. When using a hazard model with these as the only covariates and allowing for time-varying effects, we produced the plots of the cumulative regression function, seen in figure 6. The top-left curve is the cumulative baseline hazard (correspondig to those with all the others covariates equal to 0), while the others are the individual cumulative effect of having at least one other focus product. From the curves, it looks like the effect are linear for both FAM and HUS, wheras the confidence interval of OLY crosses 0 for most time interval, suggesting insignificance of influence on duration.

Combining these curves with the formal KS and CvM tests mentioned in 3.2,

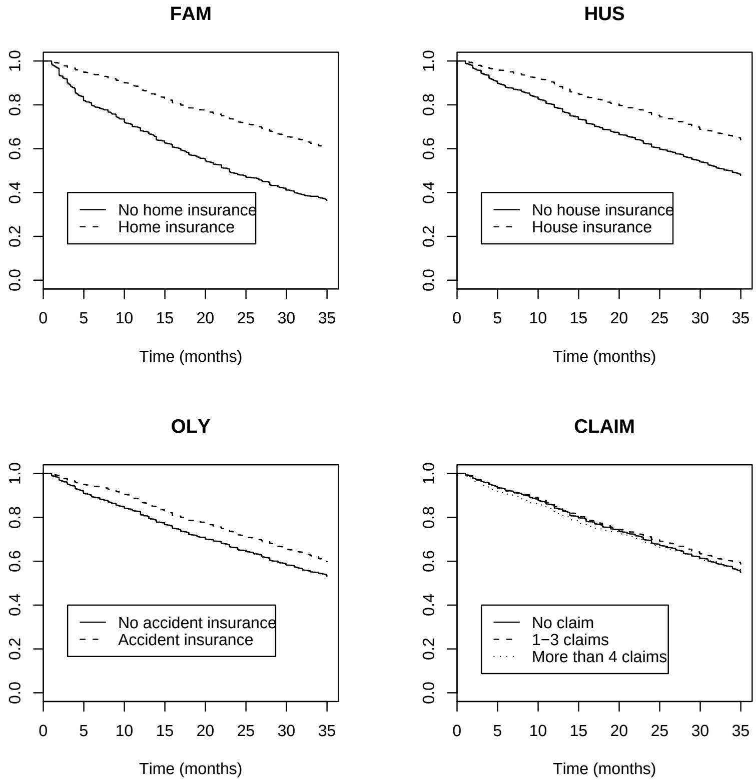


Figure 5: *Kaplain-Meier estimates on other focus products (FAM, HUS, OLY) and claim groups.*

tests of time invariant effects, suggest that the effect of FAM (i.e of having an home insurance at t_0) would vary with time - a result not easily interpreted.

```
Multiplicative Hazard Model

Test for nonparametric terms

Test for non-significant effects
            Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                       72.30                    0
fam                               10.90                    0
hus                               14.80                    0
oly                                4.62                    0

Test for time invariant effects
            Kolmogorov-Smirnov test p-value H_0: B(t)=b t
(Intercept)                 2.08                 0.128
fam                         3.58                 0.023
hus                         1.50                 0.381
```

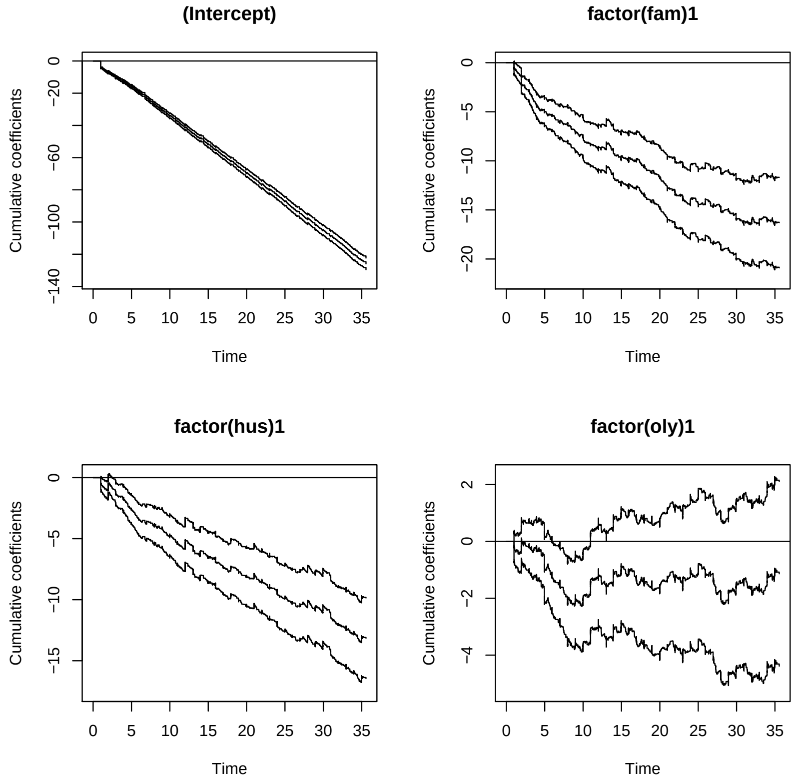


Figure 6: *Estimated cumulative regression functions with 95% pointwise confidence interval.*

```
oly                              1.99                   0.111
            Cramer von Mises test p-value H_0: B(t)=b t
(Intercept)                    50.6                   0.056
fam                           208.0                   0.001
hus                            19.0                   0.281
oly                            27.1                   0.172
```

A consequence of this, is that FAM violates the Cox model assumption and need some more care. A first method was to interact it with HUS, thus replacing the main effects of FAM and HUS with a combination, FAMHUS, of them: 1 if customer has at least one home or house insurance, 0 otherwise. Doing this results in the same conclusion, that there effect of FAMHUS is varying over time, which can be seen in the table below. The result was obtained by using the **cox.zph** function in R, which checks for devations from specified model, when modelling hazard as a time invariant Cox model with FAMHUS and OLY as only covariates. The underlying Cox model suggest that FAMHUS is significant, while OLY is not, but that both violates the Cox assumption (eqiuvalent to small p-values in the test for model violation). A incorrect interpretation here would thus be that, among motor policy holders, having a home or a house insurance in addition would decrease the churn risk by 0.5 ($= \exp(-0.694)$).

```
                   coef exp(coef) se(coef)      z    p
factor(fam_hus)1 -0.6939     0.500   0.0601 -11.55 0.00
factor(oly)1     -0.0635     0.938   0.0466  -1.36 0.17

Likelihood ratio test=149  on 2 df, p=0  n= 5000

Test for model violation:
                   rho chisq        p
factor(fam_hus)1 0.0897 17.91 2.31e-05
factor(oly)1     0.0389  3.39 6.54e-02
GLOBAL               NA 32.17 1.03e-07
```

In order to handle this time varying effect we could either create a time-dependet version of FAMHUS, which takes on different values for different, non-overlaping, time periods. Alternatively we stratify on this covariate, allowing seperate basline hazard for the two group of FAMHUS. Since we are mainly interested in applying the model to prediction, we choosed the latter.

As showned above, it seemed to be no difference between the claim groups. Indeeed, it might be difficult to show the impact of prior claims, without any further knowledge on the matter or size of the claims. This is why we included claim ratio, CLAIMPCT, which compares the claims expenses with earned premiums, in order for some measure on the profitability of the customer. If this is large, the customer have had some claims that are big in relation to their premiums during a preceding 4 year period. Due to the its skew distribution- ranging from 0 to ca 4.5 - it was reasonable to try some classfication of it. Based on a quantile distribution, we set break points at 0.025, and 0.33, creating a three level variable labelled as 'low', 'middle' and 'high' claims ratio. A stratified log-rank test, when controlling for the two stratification variables EXPMONTH and FAMHUS, is seen in table below. There is a significant difference between the groups. However, when investigating

Class, interval	n	Observed	Expected
Low, [0,0.025)	2463	1064	1127
Middle, [0.025,0.33)	1235	510	516
High, [0.33,4.5)	1302	596	526
	Chisq=13.2	on 2 df,	p=0.001

Table 5: *Stratified log-rank test for claims ratio classes*

further, the difference between the 'low' and 'middle' classes is very little, making them indistinguishable. Thus a new classification emerged by joining 'low' with 'middle' and label it is as 'low'. An updated stratified log-rank test, with only two levels, gave a chi-squared test statistic equal to 12.4 on 1 df.

To summarize, our data set before next step includes the covariates:

Index	Covariate	Description
1	ANC	Survival time, calender time (in months)
2	D	Censoring variable, 1 if event, 0 otherwise
3	TYPE	Churning cause, I.-V. if D=1, 0. otherwise
4	EXPMONTH	Expiration month on primary policy
5	ANCPRE	Duration (in months) prior to study
6	AGEGROUP	0:[20,40), 1:[40,55), 2:[55,70), 3:[70,90)
7	GROUP	1 if customer has membership in some organisation
8	FORSAAR	Total exposure over preceding 4 year period
9	CLAIMPCTG	Claim ratio groups, 1: [0,0.33), 2: [0.33,4.5)
10	FAMHUS	1 if at least 1 home or house insurance, 0 otherwise

Here we use EXPMONTH and FAMHUS as stratification variables, resulting in (12*2=) 24 baseline hazards. AGEGROUP is a 4-level covariate and is broken into binary variables AGEGROUP1 (=1 if AGEGROUP=1, 0 otherwise), AGEGROUP2 and AGEGROUP3.

5.2 Model selection and survival prediction

Before presenting final model, we check for interaction terms between the covariates in 5.1. This is done by performing ANOVA on the models and including significant interactions terms, one at the time. Let M denote the main effect model from last section and consider extension of the model of the sort $M + c_i \times c_j$, with interaction between covariates c_i and c_j for index $i, j = 5, \ldots, 9$.

When calculating the corresponding p-values for the ANOVA tables we found no significant interactions. That is, when controlling for the covariates, the trends of interaction we saw earlier in the univariate analysis disappeared.

This means that we have only main effects in the Cox model for the last step. For checking for any violation of the Cox model assumption, we performed similar test as above. As seen below, there is no individual nor global evidence that the covariates would not enter the model linearily. From the parameter estimates one concludes e.g. that the risk of churning among agegroup 3 (55-70 years) is only 66% of that of agegroup 1 (20-39 years), while controlling for everything else. Similarily, having had a high claims ratio would increase the risk of 25%.

```
Test for model violation:
                         rho   chisq      p
anc_pre             -0.00238 0.01243 0.9112
forsaar             -0.03668 2.85900 0.0909
factor(agegroup)1    0.01910 0.78932 0.3743
factor(agegroup)2    0.00855 0.15842 0.6906
factor(agegroup)3    0.01179 0.28772 0.5917
factor(grupp)1       0.00118 0.00289 0.9571
factor(skadpctg2)2 -0.04159 3.75730 0.0526
```

```
GLOBAL                    NA 8.52008 0.2890
```

```
Multivariate Cox model:
                        coef exp(coef) se(coef)     z       p
anc_pre            -0.000934     0.999 0.000224 -4.17 3.0e-05
forsaar            -0.021972     0.978 0.003910 -5.62 1.9e-08
factor(agegroup)1  -0.301842     0.739 0.061858 -4.88 1.1e-06
factor(agegroup)2  -0.413593     0.661 0.063047 -6.56 5.4e-11
factor(agegroup)3  -0.218617     0.804 0.074286 -2.94 3.3e-03
factor(grupp)1     -0.225021     0.798 0.048240 -4.66 3.1e-06
factor(skadpctg)2   0.226580     1.254 0.049144  4.61 4.0e-06
```

In the above table, the second column - exp(coef) - is the hazard ratio while controlling for everything else. z^2 is the Wald statistic and p is the corresponding p-value.

Because of the rare events of some of the causes, it is reasonable to regroup the risk classes, as discussed in 2.4. This will generally depend on the the product at hand, and for motor policies we combine riks classes I. with IV., and II with V, and let III be unchanged, in order to make inference.

Then we stack the data as described in 3.3. With the new classification of risks, we made 3 copies of the data yielding ($5 \times 3 =$) 15 covariates. By appliyng the model selection steps in the same section, we arrived at a 14 covariates model which we tested equality across risk strata. The test is a linear hypothesis on the relationship between the cause-specific covariate and conclusions are made from considering a Wald statistic.

```
Linear Hypotheses Testing Results

                           Wald
Label                Chi-Square      DF    Pr > ChiSq

EqualAnc_pre             5.8064       1        0.0160
EqualForsaar             1.7973       1        0.1800
EqualAgegroup1           0.7199       2        0.6977
EqualAgegroup2          13.0962       2        0.0014
EqualAgegroup3           4.2857       1        0.0384
EqualClaimpctg           2.4885       2        0.2882
EqualGroup               0.6056       2        0.7387
```

From the hypothesis testing, there seem to be equality among the covariates FORSAAR, AGEGROUP1, AGEGROUP3, CLAIMPCTG and GROUP. By considering the individual cause-specific (partial maximum likelihood) estimates among these covariates we get an idea which might be equal and can test if the differences are equal on a 5% significance level between all causes of just two of them.

The result from these test along with the estimated hazard ratio are seen below.

Variable	Hazard Ratio
ANC_PRE_13	0.999
ANC_PRE_2	0.996
FORSAAR_1	0.985
FORSAAR_23	0.954
AGEGROUP1_123	0.754
AGEGROUP2_1	0.725
AGEGROUP2_2	0.452
AGEGROUP2_3	0.568
AGEGROUP3_1	0.535
AGEGROUP3_2	0.143
AGEGROUP3_3	1.667
CLAIMPCTG_13	1.151
CLAIMPCTG_2	2.193
GROUP_123	0.806

The subindex indicates which effects are equal across causes. E.g ANCPRE13 indicates that ANCPRE has a similar effect on cause 1 as well 3 and that the risk of churning of cause 1 or 3 is decreasing as this covariate grows. The corresponding risk for cause 2 decreases faster.

From the output we also conclude that when comparing age group 0 (20-39 years) with age group 1 (40-54 years), the relative risk among the latter is only 75% of that among the former group and that this holds for all causes. This is not the case for the other two age groups. Especially, we notice that the relative risk of churning of cause 3 is almost 67% higher among age group 3 (70+ years) compared to age group 0. This is however expected, since it more likely that someone of age 70 or more doesn't need a car and hence no motor insurance. Cause 2 (internal churning) is more likely for younger customer.

Finally, when calculating the cumulative incidence functions, as described in 2.5.4. This gives us a tool for calculating the estimated survival probabilty given covariates. Here we apply the calculations to three customers in order to demonstrate it and we let only the age vary between them, letting everything else be the same. In this case we let each customer have had a motor policy for 2 years and had (on average) 1 policy per year and no claims (i.e CLAIMPCTG=1) and let the customer be member in some partner. Customer 0 belong to age group 0 (20-39), customer 1 to group 1 (40-55), and so forth. P_{01} is the estimated probabilty of churning of cause 1, and similarily for P_{02} and P_{03}, while $P_{00} = 1 - \sum_{j=1}^{3} P_{0j}$ is the overall survival probability.

Customer	Time (months)	P_{01}	P_{02}	P_{03}	P_{00}
0	0	0.000	0.000	0.000	1.000
0	12	0.106	0.004	0.055	0.835
0	24	0.217	0.007	0.106	0.669
0	36	0.298	0.010	0.162	0.530
1	0	0.000	0.000	0.000	1.000
1	12	0.083	0.003	0.042	0.872
1	24	0.174	0.006	0.084	0.736
1	36	0.243	0.008	0.133	0.616
2	0	0.000	0.000	0.000	1.000
2	12	0.081	0.002	0.034	0.883
2	24	0.171	0.003	0.067	0.757
2	36	0.241	0.005	0.107	0.647
3	0	0.000	0.000	0.000	1.000
3	12	0.057	0.001	0.0943	0.847
3	24	0.118	0.001	0.184	0.697
3	36	0.163	0.001	0.285	0.550

Table 6: *Cumulative incidenct function calculated at different time points for different age groups while controlling for other effects.*

What can be said from this simple example is that the most loyal customer are in the age interval 40-70, with a overall survival probability of up to ca 10% higher than those outside side interval. One can also see that churning of cause 3 (need of insurance coverage ceased) is more likely to occur among older customers, while churning of cause 1 (tranferring to competitor) is more common among younger customer, which is plausible. Churning due to cause 2 (internal churning) is more rare than the other among motor customers, but is again more likely to happen among younger customers.

Furthermore, the "rank" of the survival groups seems to be preserved over time, i.e the annual rate of churned customers seems to be about the same within groups, with some reservation for churns due to cause 2 (internal churns), which is expected, since these kinds of cancellations should tend to decrease over time.

6 Conclusions and Discussions

In this thesis, we have considered possible prediction methods for customer duration in a non-life insurance industry on a product-customer level. While concentrating on motor policy holders, the same methods could be applied to other products.

Given some data we were able to calculate the changes in the survival function, especially in a competing risk setting.

Overall, a customer is more likely to stay, depending on

1. how long the policy have been active
2. how old the customer is
3. how many policies the customer have had prior
4. how large claims the customer have had
5. if the customer has a membership with a partner

Besides these factor, we also found that structure of the model depended on renewal period and if the customer had home or house insurance, which we included in the baseline hazard.

Among these factors, we found evidence of different effects on different causes. E.g. having had an motor policy for a long time, implies a smaller risk of churning due to internal causes (that is, when the churn is initiated by the insurance company) than that of other causes. This is logical, since most internal churns on motor policies are likely to occur relatively near signing date and a long relationship would have removed most of these.

We also found that customers of age over 70 years, are 67% more likely to churn due to ceased risk (e.g. sold the car) but 50% less as likely to churn due to transferring to a competitor, compared to a customers of age 20-40. Again, this is what we can expect, since older people drive less and typically are not as price sensitive.

There might, of course, be additional information that influence the duration which might be investigated in the future.

6.1 Comments on further analysis

We considered a short study period in order for sharper predictions under the assumption that the complexity of the real world wasn't easy to capsulate, making fixed covariate models somewhat unrealistic for long-time prediction. This would suggest that it might be useful to consider time-varying, dynamic models (e.g Aalen, Aalen-Cox models) to investigate the influence of events (such as claims or move or churning of other products). It would also be interesting to consider a longer study period in order to find those customers that are truly loyal and profitable for a company, which would require a different study design than that used in here, e.g following customers from time first entering the company up to churning.

One important thing omitted here, was the model assesment and prediction capability. With survival data of this sort it could be dealt with by extended Brier scores to see where the model predicts well, while allowing for features such as censored data. The main object was however not finding an optimal model in a prediction application, but rather to suggesting a method for handeling and measuring customer duration. And in my opinion it is difficult to apply this sort of prediction algorithms to quantify customer's churning behaviours, since there is too much individual "noise" that could lead to a churn. However, measuring it and analysing it is still very important since it directly effects the profitability of the portfolio and on product level, duration might be a more manageble concept. It could then used to see where customer loyalty programs actually have an effect or where programs shold be started in order to create loyalty.

6.2 Comments on data

An important part of data in this study was that of churning causes. One might argue that this validity is difficult to assume and to investigate, but the manuel process of gathering this sort of information is accepted within the company although the precision might be questioned at times. The classification of causes was thus necessary, in an attempt to further controll the precision of causes.

In 2.4 relationships were decided to belong to different risk classes in case of cancellation. For ties (on churning dates) this meant that one class had to represent the cause of cancellation above others. For the rest, the cause of the last churned policy represented the given cause.

It is, however, likely that for a customer holding several product-specific policies that there is a more complex course of events yielding this final churn, i.e. the final churn is a result of previous churns and the timing of those. (Here, we disregard the cause of any previous, recent or not, churns.)

Thus, there are some uncertainties in defining a unique cause of an ended (product specific) relationship, since the cause of the final churn might be dependent on previous and not representative for the relationship as a whole.

But again, the classification of causes treats the causes as either initiated by the company or the customer and this level there are no ambiguities.

References

[1] Aalen, O. O. (1975). Statistical Inference for a Family of Counting Processes. Ph.D. Dissertation, University of California, Berkely.

[2] Aalen, O. O., Borgan Ø. and Gjessing, H. K. (2008) *Survival and Event History Analysis. A Process Point of View*, Springer, New York

[3] Allison, P. D. (1995). *Survival Analysis using the SAS System: A Practical Guide*, SAS Institute Inc.

[4] Anderson, P. K., Borgan O., Gill R. D. and Keiding, N. (1993) *Statistical Models Based on Couting Processes*. Springer, New York

[5] Beerli, A., Martín, J.D. and Quintana, A. (2004). A model of customer loyalty in the retail banking market. *European Journal of Marketing, Vol. 38, No. 1/2, pp. 253-275*

[6] Berry, M. and Linoff, G. (2004). *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*, Wiley Publishing, Inc.

[7] Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society Series B, Vol. 34, pp. 187-220*

[8] Crowder, M.J. (2001). *Classical competing risks*. Chapman/Hall.

[9] Dick, A. S. and Basu, K. (1994). Customer loyalty: towards an integrated conceptual framework. *Journal of Academy of Marketing Sience, Vol. 22, No. 2, pp. 99-113.*

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

[11] Guillen, M., Nielsen, J.P., Scheike, T., and Perez-Marin, A.M. (2006). Time-variying effetcs when analysing customer lifetime duration: application to the insurance market. *Institut de Recerca en Economica Aplicada*, 2006/4.

[12] Hastie, T., Tibshirani, R., and Friedman, J. (2001) *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, New York

[13] Hosmer, D. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modelling of Time to Event Data*. Wiley-Interscience

[14] Klebaner, F. C. (2005). *Introduction to Stochastic Calculus with Applications*. Imperial College Press, London

[15] Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis - Techniques for Censored and Truncated Data*. Springer, New York.

[16] Lu, J. (2002). Predicting Customer Churn in the Telecommunications Industry - An Application of Survival Analysis Modeling using SAS. *SAS User Group International Online Proceedings*, Paper No. 114-27.

[17] Martinussen, T., Scheike T. H. and Skovgaard I. M. (2002). Efficient Estimation of Fixed and Time-varying Covariate Effects in Multiplicative Intensity Models. *Scandinavian Journal of Statistics, Vol. 29, No. 1, pp. 57-74.*

[18] Rosthøj, S., Andersen, P. K. and Abildstrom, S. Z. (2004). SAS macros for estimation of the cumulative incidence functions based on a Cox regression model for competing risks survival data. *Computer methods and Programs in Biomedicine, Vol. 74, No. 1, pp. 69-75.*

[19] Scheike, T. H. and Martinussen, T. (2006) *Dynamic Regression Models for Survival Data*. Springer, New York.

[20] Shtatland, E. S., Kleinman , K. and Cain, E. M. (2005). Model Building in PROC PHREG with automatic variable selection and information criteria. *SAS Users Global Forum 2005*, Paper No. 206-30.

[21] Therneau, T. and Grambsch, P. (2000). *Modeling survival data: extending the Cox model*. Springer, New York.

[22] Van den Poel, D. and Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157, pp. 196-217*