Charles University in Prague
Faculty of Mathematics and Physics

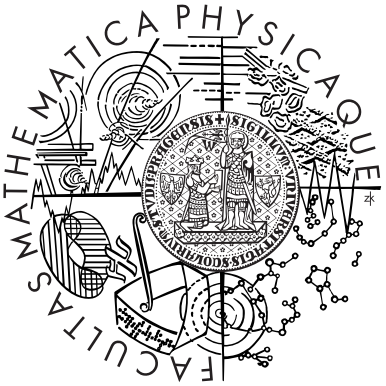
Seminar on Modelling in Economics

Survival analysis in credit scoring

Authors:
Jaroslav Pazdera
Michal Rychnovský
Petr Zahradník

email:
pazdera@matfyz.cz
michal@rychnovsky.cz
petr@anoel.eu

Supervised by:
Ing. Mgr. Jan Polívka, PhD.



Keywords
Cox model
AFT model
Survival analysis in SAS
Hazard function

In Prague, February 1, 2009

Contents

Chapter 1

Introduction

Survival analysis provides us with tools to model a time to death of an organism or a failure of a mechanical system. The concepts of survival analysis can be successfully used in many different situations, e.g. observing time until formal convicts, after having been released, commits a crime again. We fit this framework into the credit risk branch, where we observe a time to default of a client.

We present a brief introduction to survival analysis with basic notation and ideas in the first chapter. In second chapter, we postulate some models used in survival analysis and we state common estimators for them. The third chapter shows an implementation of the developed theory on data from banking sector on two specific examples using SAS 9.1. In appendix we overview the outputs for different bank product types.

We often speak about clients instead of general observations, since our data are from the banking sector and as such are specific.

1.1 Survival analysis

We consider methods for the analysis of data which model the time when a specific event occurs. In finance, we concentrate on an event called *default*, a term usually used in connection with violation of debt contract conditions, e.g. lack of will or disability to pay the debt back. Exact definition of default will be specified. Because we use survival analysis concepts, we assume that each client *will* face default at some point in the future, possibly very distant. The default may occur long after the clients' death or withdrawal from the bank, hence the nature of our approach implies heavy censoring, the term to be properly explained in section 1.2.

Let X be a nonnegative random variable representing the time of default of a client, suppose that X has a distribution function F and a density function f. The distribution of X can be specified in many ways of which authors use mainly a *hazard function*, conventionally denoted by $\lambda(\cdot)$, which is defined as

$$\lambda(t) = \lim_{h \to 0+} \frac{1}{h} P(t \leq X < t + h | X \geq t). \tag{1.1}$$

A hazard function fully specifies the distribution of X. If we define the *survival function* S(t)

by $S(t) = 1 - F(t)$, we can rewrite the hazard function as

$$\lambda(t) = \lim_{h \to 0+} \frac{F(t+h) - F(t)}{h} \frac{1}{S(t)} = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \log S(t). \tag{1.2}$$

This relation between the hazard and survival functions can also be expressed in terms of the hazard function as

$$S(t) = 1 - F(t) = \exp\left[- \int_0^t \lambda(s) \mathrm{d}s \right]. \tag{1.3}$$

For easier notation at this point, it is also useful to define the *cumulative hazard function*, denoted by $\Lambda(\cdot)$, as

$$\Lambda(t) = \int_0^t \lambda(u) \mathrm{d}u = -\log S(t).$$

We often speak about *being at risk* at time t. This implies that a client had not defaulted before time t and we were able to track him in a following period.

1.2 Censored data

We observe data from n clients. We do not usually have complete information about our client -whether he had defaulted or not- simply because we can only observe him during a fixed time interval of length T. During this interval there are three possibilities of a client status to be observed: default at time X, no default until time T or a customer leaving the survey at time C before the final status could have been obtained.

In the sense of Reisnerová (2004), the authors consider a so-called *right censoring*. Let X_i denote the time of default for each client, where X_i for $i = 1, \ldots, n$ are *iid* random variables with density f, distribution function F and other characteristics same as in section 1.1. By C_i we denote the time of censoring, where C_i, for $i = 1 \ldots, n$, are again *iid* random variables. We also observe if data are censored or not. We denote this by δ_i for each client, where $\delta_i = 1$ means that an observation of the i-th client is not censored and $\delta_i = 0$ denotes censoring. In principle, for each client we observe (T_i^*, δ_i) where, $T_i^* = \min(X_i, C_i)$.

The likelihood function for random variables X_i can be formulated as

$$L = \prod_{i=1}^{n} f(T_i^*)^{\delta_i} P(X_i > T_i^*)^{1-\delta_i} = \prod_{i=1}^{n} f(T_i^*)^{\delta_i} (1 - F(T_i^*))^{1-\delta_i}. \tag{1.4}$$

In terms of a hazard function we can rewrite (1.4) as

$$L = \prod_{i=1}^{n} \lambda(T_i^*)^{\delta_i} (1 - F(T_i^*)) = \prod_{i=1}^{n} \lambda(T_i^*)^{\delta_i} \exp\left[- \int_0^{T_i^*} \lambda(u) du \right],$$

where we have used (1.2) and (1.3). Taking right censoring into account, let T denote the longest observation period. If we assume equidistant observations in a discrete time of a unit length, i.e. $T_i^* \in \{1, \ldots, T\}$, then the risk function is constant on these time intervals. Therefore, we can formulate the likelihood function as

$$L = \prod_{i=1}^{n} \lambda(T_i^*)^{\delta_i} \exp\left[- \int_{s=0}^{T_i^*} \lambda(u) \mathrm{d}u \right] = \prod_{i=1}^{n} \prod_{t=1}^{T} \lambda(t)^{\mathrm{d}N_i(t)} \exp\left[- \int_{s=0}^{T} \lambda(u) Y_i(u) \mathrm{d}u \right], \tag{1.5}$$

where $\mathrm{d}N_i(t)$ indicates the increment of $N_i(t)$ in the time $(t-1, t\rangle$, and $N_i(t)$ takes value 0 until default of i-th client is observed, i.e. $N_i(t) = I(T_i^* < t, \delta_i = 1)$, and $Y_i(u)$, $Y_i(u) = I(T_i^* \geq u)$, indicates if i-th client is at risk at time u.

Chapter 2

Regression Models

There are several approaches to obtain the survival function from data. The most elementary and straight forward way is to calculate the survival function using the Kaplan-Meier estimator. Here we use slightly more advanced models based on an a priori knowledge of the form of a hazard function and clients characteristics as covariates.

In the following, Z_i denotes a vector of p covariates for the i-th client and β denotes the p- dimensional vector of unknown parameters.

2.1 Cox model

Suppose that covariates have a multiplicative effect on the hazard, specifically we speak about *proportional hazards* models. These models are often called Cox models, inspired by Cox (1972). He assumed that the hazard function has a form

$$\lambda(t; Z_i) = \lambda_0(t) \exp(Z_i'\beta), \tag{2.1}$$

where β is a $p \times 1$ vector of unknown parameters, Z_i are given covariates and $\lambda_0(t)$ is a *baseline* hazard function. We can interpret a baseline hazard as a hazard when all covariates are set to zero. In Cox model, the ratio of two clients obviously does not depend on the baseline function. If we consider clients with covariates Z_1 and Z_2, the proportion of their hazard functions

$$\frac{\lambda(t; Z_1)}{\lambda(t; Z_2)} = \frac{\exp(Z_1'\beta)}{\exp(Z_2'\beta)},$$

is a function of their characteristics, which does not depend on the parameter t.

Moreover, as was proposed by Cox (1972, 1975), we can extend the model and suppose that covariates are time dependent. Such a model

$$\lambda(t; Z_i) = \lambda_0(t) \exp[Z_i(t)'\beta] \tag{2.2}$$

is a straightforward generalization of (2.1). In this case one can no longer speak about proportional hazards models. In literature, the model (2.2) is often mentioned as a *relative risk model*. We will always distinguish between time dependent and time independent Cox models.

2.1.1 Likelihood function

We derive the likelihood function for the more general time dependent model only: Assume that we can observe clients in unit intervals, it follows from (1.5) that the likelihood function for Cox model (2.2) is

$$L = \prod_{i=1}^{n}\prod_{t=1}^{T}\left[\lambda_0(t)\exp(Z_i(t)'\beta)\right]^{\mathrm{d}N_i(t)}\exp\left[-\int_0^T \lambda_0(u)\exp(Z_i(u)'\beta)Y_i(u)\mathrm{d}u\right],$$

where $\mathrm{d}N_i(t)$ and $Y_i(u)$ are the same as in section 1.2. This is a fully parametric model and therefore every parameter can be estimated from maximizing the logarithm of a likelihood function. Differentiating $\log L$ with respect to the baseline function, for a fixed time t, we get

$$\frac{\partial}{\partial\lambda_0(t)}\log L = \sum_{i=1}^{n}\mathrm{d}N_i(t) - \sum_{i=1}^{n}\lambda_0(t)\exp\big(Z_i(t)'\beta\big)Y_i(t),$$

where we have used that $\lambda_0(t)$ is piecewise constant. Therefore the maximum likelihood estimator for a baseline function has the form:

$$\hat{\lambda}_0(t) = \frac{\sum_{j=1}^{n}\mathrm{d}N_j(t)}{\sum_{i=1}^{n}Y_i(t)\exp\big(Z_i'(t)\beta\big)}.$$

Such an estimator of a baseline hazard function is known as the *Breslow-Crowley* estimator.

2.1.2 Partial likelihood function

To estimate a parameter $\beta \in \mathbb{R}^p$ Cox (1972) suggested a partial likelihood approach. Consider a Cox model

$$\lambda(t, Z_i) = \lambda_0(t)\exp(Z_i(t)'\beta), \tag{2.3}$$

where $Z_i(t) = [Z_i^1(t), \ldots, Z_i^p(t)]'$ is a vector of given time-dependent covariates, $\lambda_0(t)$ is a baseline hazard function corresponding to the case where covariates are set to zero, and β is a vector of regression parameters. The corresponding survival function is

$$S(t, Z_i) = P(T > t|Z_i) = \exp\left(-\int_0^t \lambda_0(u)\exp[Z_i(u)'\beta]\mathrm{d}u\right).$$

The primary method of analysis proposed by Cox is the "conditional likelihood". Let us ignore ties in the model for a moment, i.e. in one moment there can not be more than one default in the model. Suppose, that we observe defaults in times $t_1 < ... < t_k$. Without derivation, which can be found in Kalbfleisch and Prentice (2002) section 4.4.2, we introduce a *partial likelihood*. A general partial likelihood can be expressed in the form of $L(\beta) = \prod_{j=1}^{k} f(a_j|a^{(j)}, \beta)$, where f denotes conditional density. The *partial likelihood* for the jth term of such product is according to Kalbfleisch and Prentice (2002) as follows:

$$L_j(\beta) = \frac{\lambda(t_j, Z_j)\mathrm{d}t_j}{\sum_{i=1}^{n}Y_i(t_j)\lambda(t_j, Z_i)\mathrm{d}t_j}, \tag{2.4}$$

where $Y_i(t_j)$ is again an indicator whether i-th client is at risk at time t_j , $\lambda(\cdot)$ is a corresponding hazard function and $\mathrm{d}t_j$ is an arbitrarily small time increment. Equations (2.3) and (2.4), taking a product over all times of default t_j, yield a partial likelihood formula

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(Z_j(t_j)'\beta)}{\sum_{i=1}^{n} Y_i(t_j)\exp(Z_i(t_j)'\beta)}. \tag{2.5}$$

Now we obtain the maximum partial likelihood estimate by solving a vector equation

$$\frac{\partial}{\partial\beta}\log L = 0.$$

An exact maximum likelihood estimator $\hat{\beta}$ is then given from an equation

$$\frac{\partial}{\partial\beta}\log L = \sum_{j=1}^{k}[Z_j(t_j) - K(\beta, t_j)] = 0, \tag{2.6}$$

where

$$K(\beta, t_j) = \frac{\sum_{i=1}^{n} Y_i(t_j)Z_i(t_j)\exp[Z_i(t_j)'\beta]}{\sum_{i=1}^{n}\exp[Z_i(t_j)'\beta]}.$$

All available implementations use Newton-Raphson iteration algorithm to solve (2.6), a starting value of $\beta_0 = 0$ usually suffices. If it does not, readily available methods such as halving the step help to converge and therefore, in the software such as SAS for example, the starting value is set to zero always. For more details and for an iteration algorithm see Kalbfleisch and Prentice (2002).

Asymptotic results in the model with absence of the ties give asymptotic normality for maximum likelihood estimator $\hat{\beta}$ with mean β and an observed covariance matrix $I^{-1}(\hat{\beta})$, where I is an observed information matrix.

2.1.3 Model fit statistics

To compare two or more alternative models we use the following criteria based on a likelihood function:

The statistic $-2\log\hat{L}$: If we insert the maximum likelihood estimator of β to the likelihood function (2.5) we get the maximized likelihood value $\hat{L}$. The transformation $-2\log\hat{L}$ is usually used to achieve a chi-square distribution under the null hypothesis (such that all the explanatory effects in the model, β_i, are zero). As $\hat{L}$ is less than one, the statistic $-2\log\hat{L}$ is positive and smaller values indicate better model.

Akaike's information criterion: This criterion improves the statistic $-2\log\hat{L}$ by penalizing the number of variables.

$$AIC = -2\log\hat{L} + \alpha q,$$

where q is the number of unknown parameters and α is a given constant, usually $\alpha \in (2, 6)$. For example, the SAS system uses $\alpha = 2$. AIC will therefore increase when unnecessary terms

are added to the model.

The statistic $-2\log\hat{L}$ depends on a number of observations in the set. It follows that these statistics are to be used only for comparing models on the same data sets. They are generally used for variable selection - manual or automatic selection methods implemented in the statistical software.

2.1.4 Handling ties

In Kalbfleisch and Prentice (2002) there are four methods to cope with tied data, all of which are implemented in SAS. The first is a discrete method introduced by Cox applying the partial likelihood argument to a logistic model. The second is an exact method, which is very computationally intensive for a larger number of ties. And then two approximative methods which we state below.

Suppose that we observe defaults in times $t_1 < ... < t_k$. Moreover suppose that at time t_j we observe d_j clients' defaults. Then the partial likelihood according to Peto (1972) and Breslow (1974) can be expressed as

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\sum_{i=1}^{d_j} Z_{j_i}(t_j)'\beta)}{\{\sum_{k=1}^{n} Y_k(t_j)\exp(Z_k(t_j)'\beta)\}^{d_j}}. \tag{2.7}$$

The alternative formula suggested by Efron (1977) is

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(\sum_{i=1}^{d_j} Z_{j_i}(t_j)'\beta)}{\prod_{r=0}^{d_j-1}\{\sum_{k=1}^{n} Y_k(t_j)\exp(Z_k(t_j)'\beta) - rA(\beta,t_j)\}}, \tag{2.8}$$

where $A(\beta,t_j)$ and $D(t_j)$ are defined as

$$A(\beta,t_j) = d_j^{-1}\sum_{l\in D(t_j)}\exp(Z_l(t_j)'\beta), \qquad D(t_j) = \{j_1,\ldots,j_{d_j}\}.$$

2.2 AFT model

The Cox model has an obvious intuitive meaning. However, the relation between covariates and failure time is not direct. A more intuitively appearing *Accelerated failure time (AFT)* model postulates a direct relationship between failure time and covariates; it specifies that the effect of a covariate acts multiplicatively on the failure time X or, similarly, additively on $Y = \log X$. Such an idea shall break down into a linear model:

$$Y = Z'\beta + \sigma\varepsilon, \tag{2.9}$$

where the 'error term' ε is supposed to have a density and a survival function, denoted $f(e)$ and $S(e)$ respectively. As usual, Z denotes a p-dimensional covariate vector and β is a corresponding vector of regression coefficients. The equation (2.9) can clearly be rewritten as

$$X\exp(-Z'\beta) = \exp(\sigma\varepsilon).$$

Hence the random variable $X \exp(-Z'\beta)$ has a certain hazard function, say $\lambda_0(.)$, which does *not* depend on the covariates Z. By substitution, it follows that the hazard function of X given the covariates, is following:

$$\lambda(x; Z) = \lambda_0 \left(x e^{-Z'\beta} \right) e^{-Z'\beta}.$$

The survival function for this model is

$$F(x; Z) = \exp \left(- \int_0^x \left(s e^{-Z'\beta} \right) e^{-Z'\beta} \mathrm{d}s \right)$$

The model can be extended in many ways and therefore successfully compete with all the advantages of the Cox relative risk model. There is already a developed theory including rank tests for parameter testing and related asymptotic results to cope with an AFT model leaving the density $f(e)$ unspecified and allowing the covariates to be time-dependent, to fully meet the generality of the Cox model, see e.g. Kalbfleisch and Prentice (2002)[pages 240-241] for a very brief discussion or herein mentioned references for thorough proofs.

To the best of authors knowledge, there are no reasonable implementations of such generalizations. Hence the authors decided to keep with the special case of specified 'error' distribution and fixed covariates. The 'error' distribution is in our computations defined to be standard normal for two reasons. The first reason is that the shape of a log-normal distribution hazard meets our empirical expectations and the second is that it is widely used in recent literature, see e.g. Andreeva (2006) or Hakim and Haddad (1999). A possible extension, not explored here, is to use a distribution with more parameters(degrees of freedom), e.g. Gamma distribution and thus make the model more general.

In our special case of the log-normal distribution the hazard function, λ_0^{LN}, can't be expressed directly in a closed form. However, it can be numerically computed as:

$$\lambda_0^{\mathrm{LN}}(x) = \frac{f^{\mathrm{LN}}(x)}{S^{\mathrm{LN}}(x)}, \tag{2.10}$$

where f^{LN} is the density of a log-normal distribution, i.e.

$$f^{\mathrm{LN}}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp \left[-\frac{\ln(x-\mu)^2}{2\sigma^2} \right],$$

and S^{LN} is its corresponding survival function, i.e.

$$S^{\mathrm{LN}}(x) = 1 - \int_0^x f^{\mathrm{LN}}(u) du.$$

2.2.1 Maximum likelihood

The log-likelihood function for the log-failure time variable Y is straight forward to write down. The error term can be expressed as $\varepsilon = \frac{1}{\sigma}(Y - Z'\beta)$. Using already familiar notation, transforming a random variable X we get from (1.4), that the log-likelihood is

$$l(y, \beta, \sigma) = \sum_{i=1}^{n} \log \left\{ \left[f \left(\frac{y_i - Z_i'\beta}{\sigma} \right) \frac{1}{\sigma} \right]^{\delta_i} \left[S \left(\frac{y_i - Z_i'\beta}{\sigma} \right) \right]^{(1-\delta_i)} \right\}, \tag{2.11}$$

where f is a density of the error term and S is its survival function. In our special case we have $f(x) = (2\pi)^{-1/2} e^{-1/2 \cdot x^2}$ and $S(x) = 1 - \Phi(x)$, where $\Phi(x)$ is the distribution function of a standard normal distribution and f its density.

The usual maximum likelihood approach hence applies, with its asymptotic results. To test the true parameter values in our model, we readily use the classic Wald, Rao and LLR tests as proposed e.g. in Anděl (2007).

Chapter 3

Data processing

We were given a data set of bank loans. Data contain client description variables – personal information about the client (sex, age, ...), a description of the loan (type, maturity, effect, ...) and information about an occurrence of default: a client definition of default, 90 days past due, 180 days past due (We often use the term "DPD" as an abbreviation hereafter). The following Table 3.1 gives the list of explanatory variables.

The list of given variables		
sex	age	marital status
education	employment status (since)	employer
housing status	repayment type	credit card holder
kind of employment	type of private phone	type of phone at work
No. of people in the household	monthly income	other income
limit of loan	distribution channel of loan	type of loan
date of the first drawing	maturity of loan	

Table 3.1: The list of explanatory variables from our data set potentially to be used for regression.

In total, we retrieved data from 19139 clients. The following Table 3.2 gives the percentage of defaults with regard to the specific default definitions.

Number of defaults for different definitions	
10.6%	Client definition
5.5%	90 days past due
4.3%	180 days past due

Table 3.2: Percentage of defaults in the given data set for different default definitions.

We decided to use SAS software for practical computations. It was a **good choice** since SAS has proven to be a very powerful tool in survival analysis[1] and is widely used in the banking branch.

We also received time dependent variables. In the early stage we considered the usage of the time dependent covariates in the model, but we did not use them for the final models.

[1] All SAS source code is available upon email request.

Our time dependent variables consisted of several behavioral indicators of a client during a fixed period. However, such information is not known at the time of establishing the loan and therefore it would cause troubles to interpret such approaches.

3.1 Preprocessing

We had to transform the given data in order to take non-linear effects of covariates into account. We proceeded in the following steps.

1. At first we cleaned the data in the sense of handling the obvious outliers, which could have been caused by data transfer errors. For example: A client with month salary 18 mil. CZK borrowing 10.000 CZK and paying installments.

2. We ran univariate statistics on each variable in order to find out possible inconsistency. An example: 5 loans were established in 1950 in contrast to the others in 1997.

3. For all nominal variables we had a number of categories. For our purpose it was handy to merge these categories so that each variable had at most 4 possible values. This was a sensitive process considering the number of defaults in each category and the general meaning of such categories. Using histograms we also categorized the ordinal variables accordingly. In the case of a variable `limit of loan` we preferred the human-like interpretation, that is why we have divided categories by 5000 CZK, i.e. < 10k, < 15k, < 20k, < 25k and > 25k.

4. We looked at all available variables and tried to omit the correlated ones, e.g. `number of people in the household` or `other income`.

5. Finally, we calculated the variables necessary for the model (time in months, indicators of default, etc.).

We also divided the data into two groups regarding the product type – the products with and without installments. The reason was our suspicion of a different clients' behavior in these two groups. The division was crucial, as products with installments in contrast to those without have a fixed termination date and it is shown to be an important factor in the data.

3.2 Cox model

Below, we graphically demonstrate the survival function and the hazard rate calculated from the data. As an example we chose the product without installments and with the default definition 180DPD. The significant variables for this model are stated in the Table 3.3 with their coefficients and an explanation of each category.

The survival function for these data is displayed in Figure 3.1. The survival function is displayed with its confidence interval on the level $\alpha = 5\%$.

The location of mass in the hazard function is an important factor. It shows us when the clients are most likely to default. That we have described by the hazard function for an 'average' client.

variable	category	$\hat{\beta}$
education	less than full secondary education	1.18102
	full secondary education (with school leaving exam)	0.50992
	higher then full secondary education	0
employment stability	employed $<$ 1 year	1.09880
	employed $>$ 1 year and $<$ 5 years or student or other	0.82072
	employed $>$ 5 years	0
repayment type	payment on account from same bank	-1.3370
	payment in cash or other	0

Table 3.3: Significant variables and the estimates of β are sorted by an increasing p-value.

If we fit a polynomial function of the order four without a constant term to the non-parametric estimation of a baseline function weighted by the number of observations, we get a smooth version of the baseline function, see the plot on the right in the Figure 3.1. The weighting is important since we do not have many observations at the tails. The Table 3.4 gives us the parameters of the used polynomial regression with their p-values respectively.

variable	coefficient estimation	p-value
t	0.00015216	$<.0001$
t^2	-0.00000884	$<.0001$
t^3	0.00000017	$<.0001$
t^4	-0.00000000	$<.0296$

Table 3.4: Coefficients of the polynomial regression of the hazard function for an average client

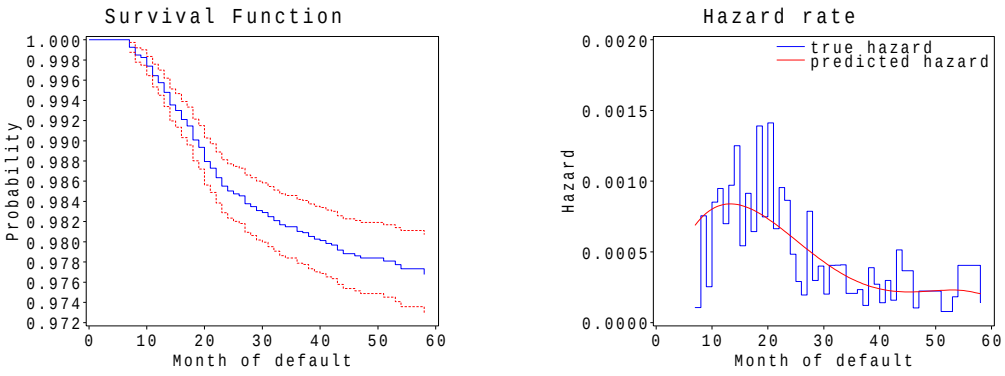


Figure 3.1: The left plot give us the survival function with confidence interval on $\alpha = 5\%$. In the right plot, there is an estimate of a hazard function for an average client and its smooth version.

Average client:
By an average client we understand the true average theoretical client, e.g. who is by 58% man and by 42% woman. To avoid any confusion we give an example of the effect on hazard rate caused by each variable. In the Table 3.5 there are averages of significant dummy variables for such a theoretical average client.

variable	category	average client
`education`	less than full secondary education	0.3919
	full secondary education (with school leaving exam)	0.3691
	higher then full secondary education	0.2390
`employment stability`	employed $<$ 1 year	0.0851
	employed $>$ 1 year and $<$ 5 years or student or other	0.3661
	employed $>$ 5 years	0.5488
`repayment type`	payment on account from same bank	0.9407
	payment in cash or other	0.0593

Table 3.5: Description of an average client in the terms of dummy variables mixture.

If we look thoroughly at the effect of an education on the hazard rate, we conclude that our computation fits the expectation that a higher education means a lower hazard rate. The hazard rate is proportional to $\exp(\hat{\beta})$. Similarly we get an expected result if we are interested in the effect of the repayment type. Clients who use a payment from the same bank account have a much smaller hazard rate. See Figure 3.2 for plotted hazard rates for each category.

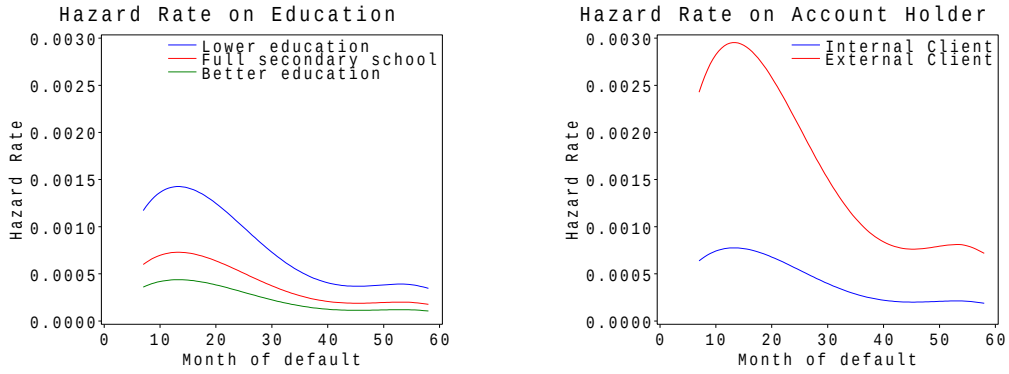


Figure 3.2: We can observe the effect on a hazard function caused by different education level of a client. The second plot confirms that internal clients, who had already hold a contract with the bank, had a lower hazard rate than the others.

3.3 AFT Model

By backward elimination, using the p-values coming out of a score test for each dummy variable, as specified in Anděl (2007), we were able to receive the significant variables for the accelerated failure time model. As we already mentioned, the baseline hazard function of

this model is fully described by one parameter, previously denoted by σ. σ was obtained by maximizing the likelihood from (2.11).

For the example described above, i.e. for a product without installments with the default definition 180DPD, it is hence straightforward to compute the hazard function for an average customer, see Figure 3.3.

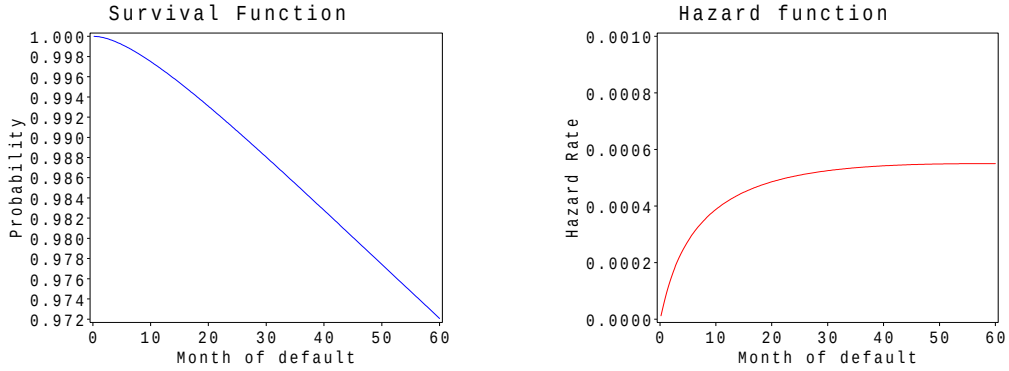


Figure 3.3: The AFT model survival function and a hazard function for an average client (compare with Figure 3.1)

Significant variables for this model are listed in the next table. The significant variables are roughly the same as in the Cox model. If we compare the estimations of parameters we can see that they don't vary too much. This fact is important for the self consistency of our models.

variable	category	$\hat{\beta}$
type of private phone	landline	0.5888
	mobile or other	0
education	less than full secondary education	-0.9722
	full secondary education (with school leaving exam)	-0.3496
	higher then full secondary education	0
employment stability	employed $<$ 1 year	-0.9614
	employed $>$ 1 year and $<$ 5 years or student or other	-0.7273
	employed $>$ 5 years	0
repayment type	payment on account from same bank	1.3823
	payment in cash or other	0

Table 3.6: Significant variables in AFT model according the increasing p-value.

A new significant variable occurs. We conclude that it is important whether the client has a landline or not. A client with a landline has a smaller hazard rate than the one without it which also fits our expectations as the clients with landlines are more 'settled down' than the clients with a mobile phone only.

Similarly to the previous sections, see the effect of education and the repayment type on the hazard rate in Figure 3.4. We get a similar result to the Cox model. Reasoning is the same: a higher education of a client means a lower hazard; an internal client means a lower

hazard. There is a huge multiplicative effect on the hazard rate and a time acceleration on the plot of the hazard rate on account holder at Figure 3.4.

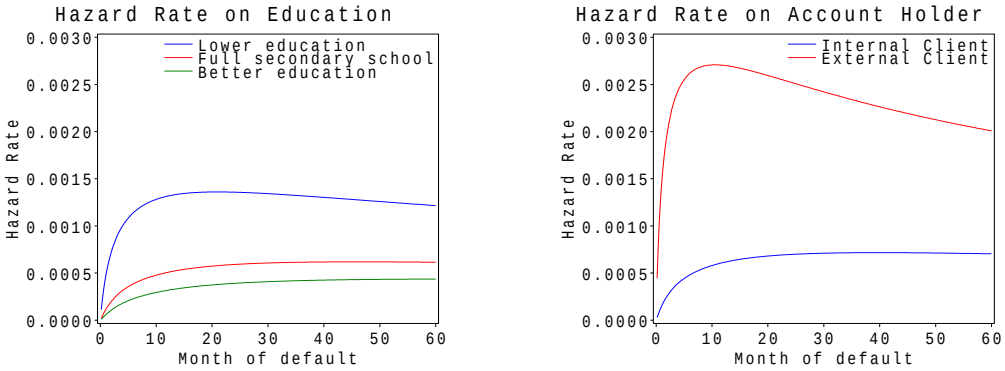


Figure 3.4: We can observe the effect on a hazard function caused by different education levels of a client on the left hand side. The second plot confirms that the internal clients of a bank have lower hazard rate than the others.

3.4 Model verification

The important criterion evaluating the models is their stability. We tested how the model and its parameters would evolve had the data set been changed.

We divided the data into two groups, so called training and testing sets. We chose randomly 2/3 of the data as a training set first. On this data, we ran our model and we got an estimate and confidence interval for β. Then we run our models on the remaining third of the data and again obtained an estimate of β. We checked whether the estimate from the testing set is within the 90% confidence interval concluded from the training set. The results were not always inside the interval which could have been caused by lack of data but at least reasonably close to.

To show an example, the Table 3.7 has the exponents of parameter estimates from a testing set, denoted by β_{test}, and the endpoints of a confidence interval gained from a training set, denoted by β_{down} and β_{up}. Because there were dummy variables used in the regression, we have one parameter less than the number of categories. We observe that the model is quite stable.

variable	category	β_{down}	β_{test}	β_{up}
education	less than full secondary education	2.675	3.025	5.605
	full secondary education (with school leaving exam)	1.099	1.537	2.477
employment stability	employed $<$ 1 year	2.181	2.958	4.229
	employed $>$ 1 year and $<$ 5 years or student or other	1.637	1.844	2.663
repayment type	payment on account from same bank	0.245	0.286	0.449

Table 3.7: Testing if an estimate of β from the test set fits the confidence interval for an estimate of β from the training set.

Chapter 4

Conclusion

We have formulated two different approaches to derive a survival function and a hazard rate. For both Cox and AFT approach we stated a likelihood function which was an essential tool for estimating unknown parameters. We applied both models on a given data set and presented satisfactory results for the product without installments with default definition 180 day past due.

The AFT model is fully described by a distribution of an error term. The AFT model will always follow this distribution, therefore once e.g. log-normal distribution is chosen, it cannot be used to describe more peaks in a hazard function. This might be considered a disadvantage, but the model is very robust and might be made more flexible by running the modeling part with different error distribution choices.

If we suspect our data set to have more than one peak in a hazard function, we should choose the Cox model. There are no assumptions or restrictions on the baseline hazard function and therefore we might be able to get more details about its true nature. In the current software the baseline is estimated using nonparametric approach, hence for easier interpretation the result usually needs to be smoothed. We have used the polynomial regression of fourth order which reveals the basic form of a hazard line and for more accuracy and better interpretation we have weighted the regression by the number of observations. However, the disadvantage of a polynomial approach is that a polynomial at the end of data 'runs away'.

For the product with installments the significant variables are different for each default definition. Usually it was the `education` and `employment type`, both described above, and often some variables from the set in the Table 4.

variable	description	impact on hazard rate
marital status	married or divorced, etc.	married have lower hazard
limit of loan	<10K, <15K, <20K, <25K, >25K	smaller limit, higher hazard, except >25K
housing status	own house or different	own house means smaller hazard
credit card holder	holding a credit card or not	credit card indicates smaller hazard

Table 4.1: An impact of other significant variables on a hazard rate.

Chapter 5

Appendix

5.1 Cox model

In the following part there is a list of graphs with hazard functions. The nonparametric plots of real hazard functions are quite rough for our data thereupon we have interpolated the curve from the polynomial regression of the fourth order.

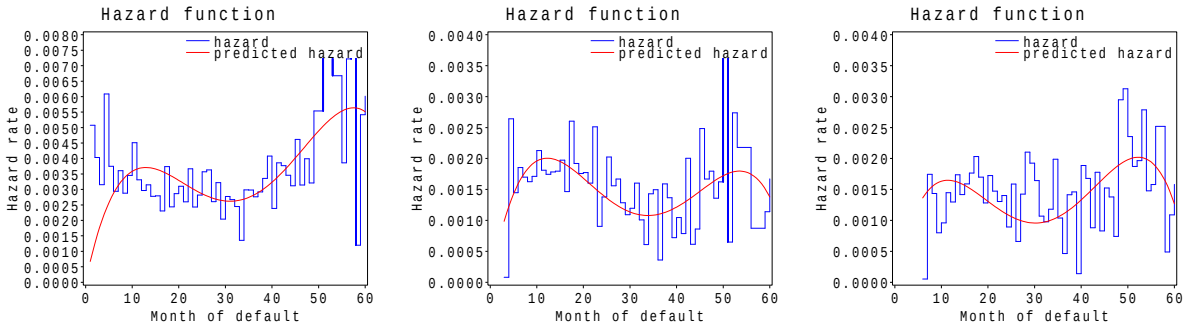


Figure 5.1: Product with installments for different definition of default. From left: hazard function of average client for client definition, 90DPD and 180DPD

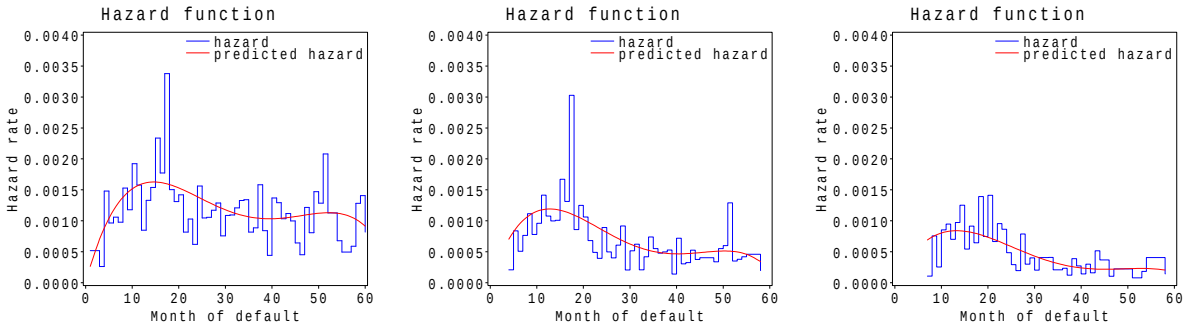


Figure 5.2: Product without installments for different definition of default. From left: hazard function of average client for client definition, 90DPD and 180DPD

5.2 AFT model

To list the results of the AFT models we don't distinguish between default types, we consider the definition 180DPD only. That is why we show two plots, one for products with installments and the second plot for products without installments. It is not necessary to show more graphical output in this paper since the graphs are similar to each other due to the limited variability of the model we used.

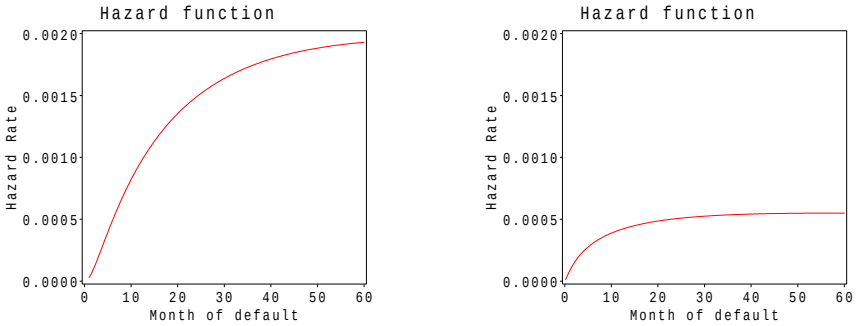


Figure 5.3: First plot is the hazard function of an average customer of product with installments, the second one is for product without installments.

Bibliography

Andreeva, G. (2006). European generic scoring models using survival analysis. *Journal of the Operational Research Society*, volume 57, no. 10, pages 1180–1187.

Anděl, J. (2007). *Základy matematické statistiky*. Matfyzpress, Praha, 2nd edition.

Breslow, N.A. (1974). Covariance analysis of censored survival data. *Biometrics*, pages 289–100.

Chava, S., Stefanescu, C., and Turnbull, S. (2008). Modeling the loss distribution. *default-risk.com*.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & HALL/CRC, 2nd edition.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, volume 34, no. 2, pages 187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, volume 62, no. 2, pages 269–276.

Dickman, P.W. (2005). Cox regression in SAS version 9. *available online at http://www.pauldickman.com*.

Fox, J. (????). Cox proportional-hazards regression for survival data, an online appendix. *An R and S-PLUS Companion to Applied Regression*. Cran.r-project.org.

Hakim, S.R. and Haddad, M. (1999). Borrower attributes and the risk of default of conventional mortgages. *Atlantic Economomic Journal*, volume 27, no. 2, pages 210–220.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure time Data*. John Wiley Sons, Inc., Hoboken, New Jersey, 2nd edition.

Peto, R. (1972). Contribution to the discussion of a paper by d.r. cox. *Journal of the Royal Statistical Society*, volume 34, pages 205–207.

Reisnerová, S. (2004). Analýza přežití a Coxův model pro diskretní čas. *Robust 2004*.

SAS Institute Inc. (2006). *SAS 9.1 documentation*. Supportsas.com.