Analysis of Life Insurance Contract Cancellations Using the Accelerated Failure Time Model

Vladimír Mucha[1] | *Bratislava University of Economics and Business, Bratislava, Slovakia*
Patrícia Ďuďák Teplanová[2] | *Bratislava University of Economics and Business, Bratislava, Slovakia*
Ján Gogola[3] | *Bratislava University of Economics and Business, Bratislava, Slovakia*
Jana Špirková[4] | *Matej Bel University in Banská Bystrica, Banská Bystrica, Slovakia*

Received 19.6.2025 (revision received 31.8.2025), **Accepted** (reviewed) 16.10.2025, **Published** 13.3.2026

Abstract

The aim of this paper is to analyse the cancellation of life insurance contracts on death using an accelerated failure time (AFT) model. The study focuses on identifying risk factors that influence the time to cancellation, with the objective of determining which to identify those insureds who cancel their policies the fastest. The analysis revealed several notable findings regarding the impact of premium payment frequency on contract cancellation. Specifically, yearly premium payments were found to extend the time to cancellation by 27% compared with monthly payments, holding all other factors constant. For contracts with monthly premiums, 10% of clients cancel within approximately 376 days, whereas for yearly premiums, the corresponding period is 476 days. Additionally, the results indicate that clients who did not conclude their contracts through the tied agent distribution channel tend to cancel their policies sooner. The AFT model was constructed using established R packages for survival analysis.

Keywords	*DOI*	*JEL code*
Survival analysis, accelerated failure time model, censoring, cancellation of an insurance contract, life insurance	*https://doi.org/10.54694/stat.2025.26*	*G22, C41*

[1] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, Bratislava University of Economics and Business, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: vladimir.mucha@euba.sk. ORCID: <https://orcid.org/0000-0001-9121-3877>

[2] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, Bratislava University of Economics and Business, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: patricia.teplanova@euba.sk.

[3] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, Bratislava University of Economics and Business, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: jan.gogola@euba.sk.

[4] Department of Quantitative Methods and Information Systems, Faculty of Economics, Matej Bel University in Banská Bystrica, Tajovského 10, 975 90 Banská Bystrica, Slovakia. E-mail: jana.spirkova@umb.sk. ORCID: <https://orcid.org/0000-0001-5864-9353>

INTRODUCTION

Modelling the time to cancellation of insurance contracts is an important analytical tool enabling the insurer to understand better the behaviour of its clients, optimise sales strategies and manage effectively its portfolio. From a long-term point of view life insurance as a product is dependent on continuation of contracts and hence identification of cancellation risk factors is one of the key aspects of actuarial analyses (Milhaud and Dutang, 2018). High cancellation rates particularly near the start of a contract can impact on the insurer's profitability (Zelinová, 2021). To analyse cancellation occurrence and the effect of explanatory variables use has also been made in actuarial practice of logistic regression, generalised linear models and penalised regression techniques (Reck, Schupp and Reuß, 2022; Kiesenbauer, 2012; Kim, 2005; Eling, and Kiesenbauer, 2014). Such methods however cannot cope with incomplete data, i.e. with censored data. The investigated event will not arise in respect of all the subjects entering the period of observation. It would not however be correct to exclude them from the analysis. Hence the term time-censorship was introduced and for such subjects we observe the censored survival time (Kleinbaum a Klein, 2012). The most common form of censorship in survival analysis is that from the right (Moore, 2016). Regression models for survival analysis differ from classical regression models in that they allow us to make use also of censored data. Parametric models are also suitable for analysis of data censored from the left, from the right as well as within an interval, as compared with non-parametric models which are only able to deal with right-censored data (Collett, 2015).

The Cox model of proportional risk and the parametric model of the accelerated failure time are important regression models in survival analysis, whereby many authors use them for example to model client survival or insurance contract cancellations. Analysis, using the Cox proportional model, of the survival of breast cancer patients depending on the therapeutic approach adopted is dealt with in Abadi etc. (2014). The patients were divided into eight groups according to age and stage of illness. For each group they applied a suitable model according to meeting the assumed proportional risks, which they then tested with the help of Schoenfeld residuals. Sheng, Qian and Ruan (2018) investigated the factors affecting the survival of heart failure patients using the Cox proportional model. Majeed (2020) in their analysis of the length in force of insurance contracts gave preference to the accelerated failure time model over the Cox model as the latter requires fulfilment of the proportional risk assumption whilst the former does not. The analysis was carried out on life insurance data from the USA, whereby the best model according to various criteria was the AFT model using a generalised gamma distribution for the time. Aziz and Razak (2019) used the Kaplan-Meier estimate and the Cox model to identify the riskiest group of life insurance clients based on Malaysian data for the years 2012–2015. In Li (2017) the author analyses survival of patients diagnosed with breast cancer using a new model of proportional risk, namely the hypertabastic model. In recent times more use is being made of joint models for longitudinal and time-to-event data in survival analysis. These models combine the ability to analyse longitudinal data and survival data and provide better results than the classical models, see Baart, Boersma and Rizopoulos, (2019). An interactive guide to carrying out a survival analysis with data on lung cancer using Python, with the aim of developing an accessible application in Streamlit, is provided in Komara and Zelinová (2024).

Machine learning methods are amongst the most innovative approaches for data analysis and are used also in survival analysis for example Štepánek et al. (2021, 2023). Implementation of the AFT model in the XGBoost library in the Python language increased the effectiveness of modelling thanks to the technique of gradient boosting (Barnwal, Cho, and Hocking, 2022). The authors Yang et al. (2021) and Ramezankhani et al. (2017) in their papers showed the use of survival trees. Kasaraneni (2024) investigated the application of advanced machine learning techniques, including deep neuron networks and Recurrent Neural Networks (RNN) to predict cancellations. Azzone et al. (2021) use the random forest method to predict cancellations of life insurance contracts. To apply these regression models in survival analysis and visualise the results use can be made of, for example, the library `survival`, `survminer`, available in R (Therneau, 2023), (Kassambara and Kosinski, 2021). This paper focuses on analysing the impact of premium payment frequency on contract cancellations, as well as examining the influence of the distribution channel. Based on the modeling of cancellation times using the AFT model, we quantify the percentage share of insurance contracts that are expected to be cancelled within the modelled time horizon.

1 METHODS OF ANALYSIS

The most often used model for survival analysis is the Cox semi-parametric regression model of proportional risks. The proportional risk model is expressed by the regression model for the risk function (Collett, 2015):

$$h(t \mid \mathbf{x}) = h_0(t)\exp(x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p) = h_0(t)\exp(\boldsymbol{\beta}\mathbf{x}), \tag{1}$$

where **x** is a vector of explanatory variables, p is their number, **β** is a vector of regression coefficients, whose elements are β_m for $m = 1, 2, ..., p$ and $h_0(t)$ is a basic risk function valid for all referential observations $\mathbf{x} = \mathbf{0}$.

The hazard ratio, HR, of the two sets of observations states how the risk function of the observations with values of the explanatory variables $\mathbf{x}_1$ differs compared with the risk function of the observations for which the explanatory variables take the values of the vector $\mathbf{x}_2$. We can express this as follows (Teplanová, 2023):

$$HR = \frac{h(t \mid \mathbf{x}_1)}{h(t \mid \mathbf{x}_2)} = \frac{h_0(t)\exp(\boldsymbol{\beta}\mathbf{x}_1)}{h_0(t)\exp(\boldsymbol{\beta}\mathbf{x}_2)} = \exp\left[\boldsymbol{\beta}(\mathbf{x}_1 - \mathbf{x}_2)\right]. \tag{2}$$

We see that the hazard ratio is independent of the time t and therefore is the same for all points of time, which is the assumption of the proportional risk model. This model not only enables, by using the hazard ratio, comparison of two groups and quantification of the risk of occurrence of the investigated event but also an analysis of the effect of risk factors on its occurrence. Given the aim of this paper we will not consider this model in detail. We will give more space to the alternative AFT model, which does not require keeping proportionality of risks (Wei, 1992; Saikia and Barman, 2017). As opposed to the Cox model of proportional risk, where the multiplicative effect of the explanatory variables is applied to the risk function, in this model it is the time to the occurrence of the given event.

1.1 The accelerated failure time model

The accelerated failure time model is therefore one of the alternatives for comparing the survival time of two or more groups of objects. Also it is a parametric model and therefore it is necessary to choose the right probability distribution for the time to the occurrence of the event. It is suitable for analysing left-censored data, right-censored data and interval-censored data as compared with non-parametric models which can only cope with right-censored data (Klein and Moeschberger, 1997). The random variable of the time to occurrence of the event in the case of the AFT model is given by the Formula:

$$T = \exp(\beta_0 + \boldsymbol{\beta}\mathbf{x} + \sigma\varepsilon), \tag{3}$$

where β_0 is the intercept, **β** is the vector of regression coefficients, **x** is the vector of explanatory variables, ε represents a random error term and σ is the standard deviation of this random element. Similarly the logarithm of the time to occurrence of the event is given by (Moore, 2016):

$$\ln T = \beta_0 + \boldsymbol{\beta}\mathbf{x} + \sigma\varepsilon. \tag{4}$$

For the survival function in context with $S_0(t) = P(T > \mathrm{t}|\mathbf{x} = \mathbf{0})$ we have:

$$S(t\,|\,\mathbf{x}) = S_0\left(t \cdot \exp\left(-\boldsymbol{\beta}\mathbf{x}\right)\right). \tag{5}$$

For the survival function in context with the survival function of the random error term $S_\varepsilon\,(t)$we have:

$$S(t\,|\,\mathbf{x}) = S_\varepsilon\left(\frac{\log\left(t\right) - \beta_0 - \boldsymbol{\beta}\mathbf{x}}{\sigma}\right). \tag{6}$$

In Table 1 we give a selection of possible distributions for the time T to occurrence of the event and for the random error term ε (Klein, 2014).

Table 1 Selected probability distributions for ***T*** and ***ε*** used in the AFT model

Distribution of T	Distribution of ε
Log-Normal	Normal(0;1)
Log-Logistic	Logistic(0;1)
Weibull	Gumbel(0;1)

Source: Own construction

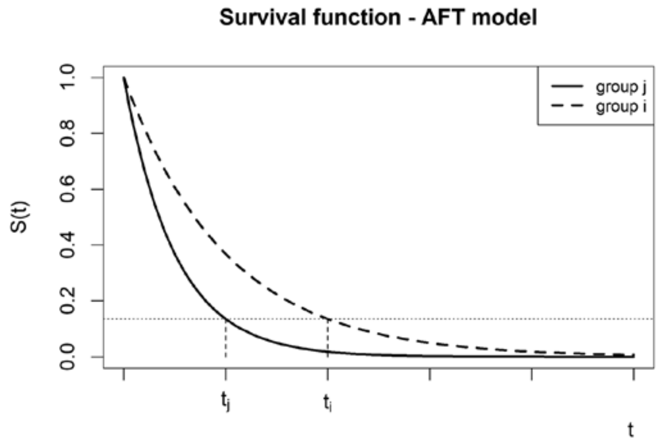
For two distinct groups i, j with differing values of covariates represented by the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, the following relationship holds between their survival functions (Collett, 2015), see Figure 1:

$$S_i(t_i\,|\,\mathbf{x}_i) = S_j(\kappa^{-1} \cdot t_i\,|\,\mathbf{x}_j) = 1 - \alpha\ , \tag{7}$$

where κ is a constant representing the acceleration factor, which can be expressed as:

$$\kappa = \exp\left[\left(\mathbf{x}_i - \mathbf{x}_j\right)\boldsymbol{\beta}\right]. \tag{8}$$

Figure 1 Survival function for groups ***i*** and ***j*** with acceleration factor ***κ*** > **1**



Source: Own construction in R

Based on Formula (7), and assuming the same value of 1 – α, the following relationship holds for the time ratio between the two groups *i*, *j*:

$$\frac{t_i}{t_j} = \kappa. \tag{9}$$

If $\kappa > 1$, the time to occurrence of the observed event for group *i* as compared with group *j* is "extended" by 100(κ – 1)%, or to put it differently for group *i* time passes κ times "more slowly" as compared with group *j*.

If $\kappa < 1$, the time to occurrence of the observed event for group *i* as compared with group *j* is "shortened" by 100(1 – κ)%, or to put it differently for group *i* time passes κ^{-1} times "faster" as compared with group *j*.

The most frequently used parametric model for describing time in the context of use of the AFT model is the Weibull distribution, whose density function $T \sim Weibull(\lambda; \gamma)$ is given by:

$$f(t) = \frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1} \cdot \ \exp\left(-\left(\frac{t}{\lambda}\right)^{\gamma}\right), t > 0. \tag{10}$$

In this case the random element takes a Gumbel distribution $\varepsilon \sim Gumbel(0;\, 1)$. Based on Formula (6), the survival function $S(t|\mathbf{x})$ of the AFT Weibull model can be expressed as:

$$S(t \mid \mathbf{x}) = \exp\left(-\left(\frac{\mathrm{t}}{\exp\left(\beta_0 + \boldsymbol{\beta}\mathbf{x}\right)}\right)^{\frac{1}{\sigma}}\right). \tag{11}$$

If we introduce the notation $\lambda = \exp(\beta_0 + \boldsymbol{\beta}\mathbf{x})$ and $\gamma = \frac{1}{\sigma}$, we obtain the survival function of the Weibull distribution (Majeed, 2020):

$$S(t \mid \mathbf{x}) = P(T > t \mid \mathbf{x}) = S(t) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\gamma}\right). \tag{12}$$

If we denote $S(t \mid \mathbf{x}) = P(T > t \mid \mathbf{x}) = 1 - \alpha$, then from the expression for the quantile function $F^{-1}(\alpha)$ of the Weibull distribution with estimated parameters λ, γ, we obtain a formula for calculating the time to the occurrence of an event based on the given value of α as:

$$t_\alpha = F^{-1}(\alpha) = \left[-\log(1-\alpha)\right]^{\frac{1}{\gamma}} \cdot \lambda, \qquad \alpha \in (0;1). \tag{13}$$

1.2 Testing the statistical significance of the variables and choice of model

This part deals with testing the statistical significance of the regression coefficients and choice of the most suitable model. If a regression coefficient appears as statistically significant, then the parameter of the explanatory variable, representing this coefficient, has a statistically significant effect on the time to occurrence of the event (Kleinbaum and Klein, 2012).

We define a null hypothesis (regression coefficient is not statistically significant) and an alternative hypothesis (regression coefficient is statistically significant) thus:

$H_0 : \beta_m = 0$,
$H_1 : \beta_m \neq 0$,
for $m = 1, 2, \ldots, p$ parameters.

We use the Wald test, where the Wald statistic Z_m^W is calculated as follows:

$$Z_j^W = \left(\frac{\hat{\beta}_m}{sd\left(\hat{\beta}_m\right)} \right)^2 . \tag{14}$$

The Wald statistic Z_m^W has an asymptotic chi-squared distribution with one degree of freedom. We reject the null hypothesis H_0 at the significance level α, if:

$$Z_m^W > \chi^2_{1-\alpha,1} , \tag{15}$$

or if the p – value is less than the chosen significance level α, whereby:

$$p - value = P\left(\chi^2 > Z_m^W\right). \tag{16}$$

For the total statistical significance of the categorical variable the Wald statistic has a chi-squared distribution with degrees of freedom equal to (number of variants – 1). To choose the most suitable model we use the Akaike information criterion (*AIC*), defined as follows:

$$AIC = -2 \ln \hat{L} + 2q, \tag{17}$$

where q represents the number of estimated parameters in the model and $\hat{L}$ is the maximized value of the likelihood function for the model. The smaller the value *AIC* the more suitable is the model. There also exist other criteria, for example Bayes information criterion *BIC*.

1.3 Verification of the AFT Weibull model using standardised residuals

We calculate the standardised residuals for the AFT Weibull model for the k-th observation in accordance with Formula (6) with the help of the Formula (Collett, 2015):

$$r_k \cong \varepsilon_k = \frac{\log\left(T_k\right) - \beta_0 - \boldsymbol{\beta}\mathbf{x}}{\sigma}. \tag{18}$$

In accordance with Formula (11) we assume that the survival function of the residuals $S_r(t)$ will be identical to the survival function of the Gumbel distribution $\varepsilon \sim Gumbel(0; 1)$, which we can write as:

$$S_r\left(t\right) \cong S_\varepsilon\left(t\right) = \exp\left(-\exp\left(-t\right)\right). \tag{19}$$

2 DATA DESCRIPTION AND MODEL BUILDING

Using actual data from insurance practice we will investigate the effect of various risk factors on the cancellation of life insurance products, whose main risk element is a benefit of freely chosen amount payable on death with the possibility of critical disease and invalidity riders.

2.1 Data description

The data looked at consists of 25 364 insurance contracts sold over a period of 10 years. The individual variables appearing in the model are described in Table 2. The explanatory variable is the time a contract is in force, i.e. the time from the date of inception to the date of cancellation, to maturity, to the ending of the risk or to the date when the contract ceases to be observed.

Let us define the variable STATE as a censoring indicator. The value "1" applies to those contracts which were cancelled during the observed period. For contracts which mature, or the risk ends, or which are still in force it takes the value "0". It should be noted that for the purpose of our analysis we include amongst cancelled contracts also those where no surrender value is paid on cancellation. In the given portfolio, more than 53.2% of the contracts were cancelled, while approximately 46.8% remained active. The number of cancellations and non-cancellations is the same. So for this data base we do not have the problem of sparse data. Clients have the possibility of including an investment element in their contract. An important factor regarding client loyalty is the care they receive from the distribution channel through which they bought the contract. This particular contract was sold through its own bank-insurance channel "Bank", insurance brokers "Broker" or tied agents "TA". As already mentioned, clients could choose to include rider benefits: invalidity (INV), critical illness (CRIT) or both combined (COMB), or choose not to include them (NR). When completing the contract the client can choose how frequently the premiums will be paid: "1" = monthly, "3" = quarterly, "6" = half-yearly and "12" = yearly.

Table 2 Description of the data base for analysing the departure of clients from the insurance company

Name of the variable	Type of variable	Values	Variable description
TIME	Continuous	---	Time in force
STATE	Category	0	Indicates if contract is cancelled or not
		1	
INVEST	Category	0	Indicates if contract contains an investment element
		1	
YEARLY_PREMIUM	Continuous	---	Amount of annual premium
AGE	Continuous	---	Age of client at inception
SEX	Category	F	Sex of insured person
		M	
DCH	Category	BANK/BROKER	Distribution channel through which the contract was sold
		TA	
RIDER	Category	NR	No rider benefits
		CRIT	Critical illness
		INV	Invalidity
		COMB	Both combined
SUM_ASSURED	Continuous	---	Contractual amount insured for each risk
FREQ	Category	1	Frequency of payment of the premium (in months)
		3	
		6	
		12	

Source: Own construction

To interpret better the results it is desirable to specify the reference categories (levels) for the categorical variables. Table 3 shows the reference level for each of the category variables.

Table 3 Reference levels

Variable	Reference level
SEX	M – males
DCH	BANK/BROKER (BB)
INVEST	0 = no investment component
RIDER	0 = without rider benefits
FREQ	1 = monthly premiums

Source: Own construction

2.2 Model building

Our aim is to estimate an AFT model which suitably describes the time to cancellation of the contracts in our data base. We will create three parametric regression models (exponential, Weibull, log-logistic). To start with it is desirable to test which variables to include in the model given their contribution to the variability of the explanatory variable. To investigate the overall statistical importance of each variable, we applied the backward elimination method. This gave for all of the three models that the best in each case was a complete model including all the variables. To choose which of the three AFT models was the most suitable we used the Akaike information criterion *AIC* and the Bayes information criterion *BCI*. Based on the results we can assert that the most suitable parametric model for modelling the time to cancellation of an insurance contract with the lowest *AIC* and *BIC* is the Weibull AFT model, Table 4.

Table 4 Test statistics for the individual models considered

MODEL	AIC	BIC
Exponential model	246 708.5	246 814.5
Weibull model	**246 535.9**	**246 649.9**
Log-logistic model	246 666.4	246 780.3

Source: Own construction

The estimation of regression coefficients and the assessment of the statistical significance of the variables using the Wald test were carried out in R. All the regression coefficients, also in terms of the variations, are statistically significant at the 0.05 significance level, apart from variation FREQ6. For interpretation purposes we have however included it in the model. To calculate the time to cancellation of the contract based on Formula (13) we determine the parameter γ from these results as follows:

$$\gamma = \frac{1}{\exp\left(Log\left(scale\right)\right)} = 0.901.$$

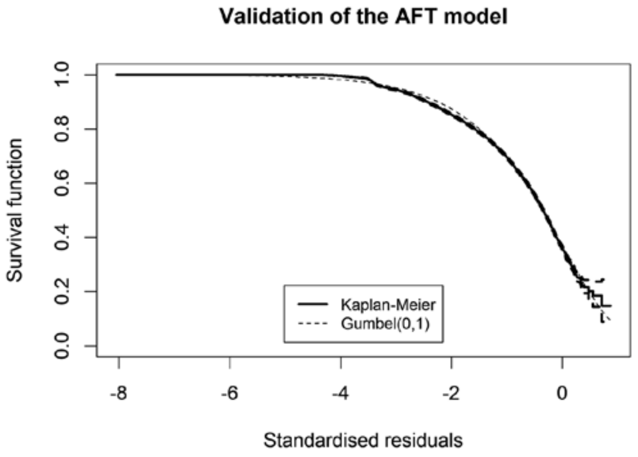
Table 5 presents the estimated regression coefficients for the AFT Weibull model, along with the corresponding acceleration factors. For categorical variables, these factors are interpreted relative to their reference category, while for continuous variables, they represent the effect of a one-unit increase, with all other conditions held constant.

Table 5 The regression coefficients and acceleration factors for the AFT Weibull model

Variables, variations	*m*	$\widehat{\beta}_m$	$\kappa = exp(\widehat{\beta}_m)$	$\kappa^{-1} = exp(-\widehat{\beta}_m)$
SEXF	1	0.085218302	1.088954761	0.918311793
YEARLY_PREM	2	–0.000319505	0.999680546	1.000319556
AGE	3	0.002121354	1.002123605	0.997880895
DCH_TA	4	0.326765333	1.386476079	0.721252978
INVEST1	5	4.536438584	93.35772169	0.010711487
RIDER_INV	6	0.828061500	2.288877448	0.436895388
RIDER_COMB	7	2.284567444	9.821436987	0.101818095
RIDER_CRIT	8	2.282594229	9.802076285	0.102019202
SUM_ASSURED	9	–0.000002320	0.999997683	1.000002317
FREQ3	10	–0.173958084	0.840332120	1.190005684
FREQ6	11	0.085279332	1.089021223	0.918255750
FREQ12	12	0.236618032	1.266957089	0.789292715

Source: Own construction

Figure 2 Validation of the AFT Weibull model using the standardised residuals



Source: Own construction in R

To validate the suitability of the estimated AFT Weibull model we use the standardised residuals. Figure 2 shows the survival function of the empirical data estimated using the Kaplan Meier estimate (Teplanová and Páleš, 2021) with the survival function of the Gumbel distribution. We can state in the context of Formula (19) that the estimated AFT Weibull model is a suitable model for describing the time to cancellation of an insurance contract.

3 RESULTS AND DISCUSSION

In this part of the paper we will interpret the obtained regression coefficients of the AFT Weibull model and analyse the influence of the chosen risk factors on the insurer's cancellation experience.

3.1 Interpretation of the regression coefficients and the acceleration factors

Based on the results shown in Table 5 we will bring out the importance and usefulness of the estimated regression coefficients for the purpose of interpreting the cancellation risk in the context of both quantitative (continuous) and qualitative (category) variables. First we will look at the effect of the sex of the insured on the modelled time to cancellation. The time ratio for the sex variable is $\kappa = 1.09$ (acceleration factor):

$$t_{females} = 1.09 \cdot t_{males}, \qquad S_{females}(t \mid \mathbf{x}) = S_{males}(0.92 \cdot t \mid \mathbf{x}),$$

which indicates that for females the time to cancellation passes 1.09 times "slower" than for males, respectively it extends it by 9% other things being equal. This means that on average males cancel their contracts earlier than females subject to the other variables remaining unchanged.

Let us demonstrate the calculation of the time ratio for example for a 20 and a 55 year old insured. Age in the model is a continuous variable and so we can use Formula (8) to quantify it:

$$\kappa = \exp\left(0.002121354 \cdot 35\right) \cong 1.08, \qquad S_{AGE=55}(t \mid \mathbf{x}) = S_{AGE=20}(0.93 \cdot t \mid \mathbf{x}).$$

With an increase in age at inception from 20 to 55 the time to cancellation extends by 8% other variable remaining unchanged.

The group who purchased their contracts through tied agents has on average a 1.39 times longer time before cancellation than the reference group who purchased through a bank or a broker. Time for this group passes 1.39 times "slower" other variables remaining unchanged:

$$t_{TA} = 1.39 \cdot t_{BB}, \qquad S_{TA}(t \mid \mathbf{x}) = S_{BB}(0.72 \cdot t \mid \mathbf{x}).$$

For insureds who had an investment element in their contract time passed 93.36 times "slower" than for those who did not, other variables remaining unchanged:

$$t_{INVEST1} = 93.36 \cdot t_{INVEST0}, \qquad S_{INVEST1}(t \mid \mathbf{x}) = S_{INVEST0}(0.01 \cdot t \mid \mathbf{x}).$$

The result could however have been affected by the fact that in the portfolio considered the number of contracts with an investment element was significantly smaller than the number without. So one can nevertheless say that in the given portfolio clients with an investment element hardly ever cancel their contracts.

Regarding frequency of payment of the premium, we see that for contracts with premiums paid once a year the time to cancellation "extends" by 27% compared with contracts where premiums are paid monthly, other variables remaining unchanged, i.e. we have:

$$t_{FREQ12} = 1.27 \cdot t_{FREQ1}, \qquad S_{FREQ12}(t \mid \mathbf{x}) = S_{FREQ1}(0.79 \cdot t \mid \mathbf{x}).$$

3.2 Modelling the time to cancellation using the AFT Weibull model

This section deals with modelling the time to cancellation using the estimated AFT Weibull model. First though we will analyse the effect of premium payment frequency on the time to cancellation for a contract on a 30 year old female, taken out through tied agents, with a yearly premium of € 200 and sum assured of € 5 000 without rider benefits and without an investment element. We model these times for different levels of the survival function values, $1 - \alpha$; that is, we estimate, with a predefined probability, the duration of the contract. To estimate the time to cancellation of an insurance contract, we use Formula (13) of the Weibull AFT model, based on which we obtained the resulting expression:

$$t = [-\log(1- \alpha)]^{1.11} \cdot \exp\,(8.026686233 + 0.085218302 \cdot \mathit{sex\_female} - 0.000319505 \cdot \mathit{yearly\ premium} + 0.002121354 \cdot \mathit{age} + 0.326765333 \cdot \mathit{DCH\ tied\ agents} + 4.536438584 \cdot \mathit{investment\ element} + 0.828061500 \cdot \mathit{invalidity} + 2.282594229 \cdot \mathit{critical\ illness} + 2.284567444 \cdot \mathit{combined\ riders} - 0.000002320 \cdot \mathit{sum\ assured} - 0.\ 173958084 \cdot \mathit{freq.quarterly} + 0.085279332 \cdot \mathit{freq.halfyearly} + 0.236618032 \cdot \mathit{frek.yearly}).$$

Table 6 shows the values rounded to two decimal places of the times to cancellation with the values of the acceleration factors for each variation of the frequency of premium payment as compared with the reference monthly frequency. The table shows the modeled times to contract cancellation only for selected values of the survival function. The emphasised figures represent the median time $t_{0.5}^{FREQf},\ f = 1,3,6,12$ to cancellation of the contract for the client with the aforementioned risk profile for the various frequencies of payment of the premium. Given these values and subject to maintaining the conditions of the Bernoulli theorem of large numbers, we can state for example that for monthly payment of premiums 50% of contracts are still active, i.e. in force, approximately after 3 041 days.

Table 6 Modeling the time to cancellation for various premium payment frequencies

$S(t)$	$t_{monthly}$	$t_{quarterly}$	$t_{halfyearly}$	t_{yearly}	$\kappa_{\frac{quarterly}{monthly}}$	$\kappa_{\frac{halfyearly}{monthly}}$	$\kappa_{\frac{yearly}{monthly}}$
1	0	0	0	0	0	0	0
0.9	375.75	315.76	409.20	476.07	0.84	1.09	1.27
0.8	864.29	726.29	941.23	1 095.02	0.84	1.09	1.27
0.7	1 454.64	1 222.38	1 584.13	1 842.96	0.84	1.09	1.27
0.6	2 167.28	1 821.23	2 360.21	2 745.85	0.84	1.09	1.27
0.5	**3 041.22**	**2 555.64**	**3 311.96**	**3 853.10**	0.84	1.09	1.27
0.4	4 145.61	3 483.69	4 514.66	5 252.32	0.84	1.09	1.27
0.3	5 613.28	4 717.02	6 112.98	7 111.78	0.84	1.09	1.27
0.2	7 747.12	6 510.15	8 436.78	9 815.27	0.84	1.09	1.27
0.1	11 528.99	9 688.18	12 555.32	14 606.74	0.84	1.09	1.27

Source: Own construction

We can write this as follows:

$$S(3041.22\,|\,\mathbf{x}) = P(T > 3041.22\,|\,\boldsymbol{x}) = 0.5, \qquad t_{0.5}^{FREQ1} = 3041.22.$$

For comparison, with the yearly payment frequency (*FREQ*12)cancellation occurs after approximately 3 853 days, which we can write as:

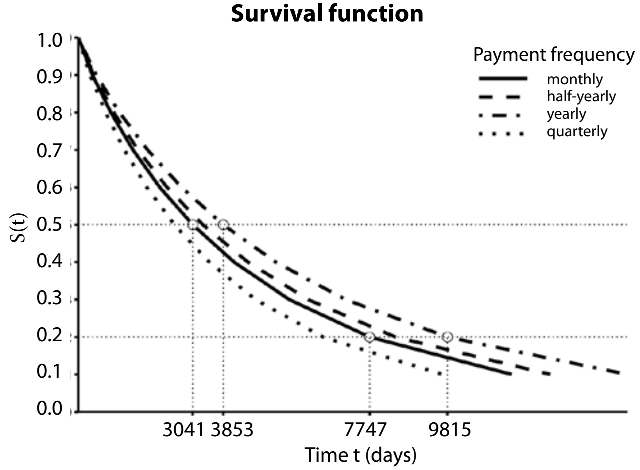
$$S(3853.10\,|\,\mathbf{x}) = P(T > 3853.10\,|\,\boldsymbol{x}) = 0.5, \qquad t_{0.5}^{FREQ12} = 3853.10.$$

Given these median times to cancellation (approximately 3 041 days for monthly payments and 3 853 for yearly payments), the acceleration factor, other things being equal, is $\kappa \cong 1.27$. So for the group of insureds with yearly payment of premiums time flows "slower" compared with the reference category (monthly payment of premiums). It is clear that the graph of the survival function is moved to the right of that for the reference group, as illustrated in Figure 3, $t_{0.5}^{FREQ12} = 1.27 \cdot t_{0.5}^{FREQ1}$. Figure 3 also shows the 80th percentiles of the time to cancellation for monthly and yearly payment frequencies, which we can write as:

$$S(7\,747.12 \mid \mathbf{x}) = P(T > 7\,747.12 \mid \boldsymbol{x}) = 0.2, \quad t_{0.80}^{FREQ1} = 7\,747.12,$$

$$S(9815.27 \mid \mathbf{x}) = P(T > 9815.27 \mid \boldsymbol{x}) = 0.2, \quad t_{0.80}^{FREQ12} = 9815.27.$$

Figure 3 Survival function for the given risk profile with various payment frequencies



Source: Own construction in R using package ggplot2 (Wickham, 2016)

In the case of clients who pay yearly premiums 80% cancel their contracts within approximately 9 815 days, which is 1.27 times later than for clients who pay monthly premiums: $t_{0.8}^{FREQ12} = 1.27 \cdot t_{0.8}^{FREQ1}$.

In the case of clients who pay monthly premiums 10% cancel within approximately 376 days ($t_{0.1}^{FREQ1} = 375.75$), whereas for yearly premium contracts it is 476 days ($t_{0.1}^{FREQ12} = 476.07$).

Once again we point out that these results apply for the risk profile of a female aged 30 who bought a contract via a tied agent with an annual premium of € 200 and an insured amount of € 5 000, without rider benefits or an investment element.

If the distribution channel is changed from a tied agent to a bank/broker (BB), with a yearly premium payment frequency and all other parameters held constant, the results are presented in Table 7. A comparison of these results with those in Table 6 shows that where a client buys the contract via a bank or broker time passes "faster". This means that they cancel their contracts earlier than those who buy via tied agents. Just for comparison 10% of bank/broker clients cancel within approximately 343 days ($t_{0.1}^{BB} = 343.3514$), whereas it is 476 days ($t_{0.1}^{TA} = 476.07$) for tied agent clients.

Table 7 Estimate of selected quartiles for a client with distribution channel BB

$S(t) = 1 - \alpha$	t_{α}^{BB}
0.9	343.4
0.5	2 779.0
0.2	7 079.1

Source: Own construction

CONCLUSION

The parametric regression AFT model allows us to model the time to the cancellation of an insurance contract, as well as to analyse how individual risk factors affect it. In this context it is also suitable for comparing client groups with different risk profiles using the acceleration factor. These calculated time ratios quantify in which of the compared groups time passes "slower" or "quicker". The quantile values of the AFT model allow us to predict the percentage of contracts, with a given risk profile, which will be cancelled up to an estimated time. After processing actual data relating to a portfolio of life insurance contracts whose main benefit was an amount payable on death, it was determined that the most suitable model for describing the time to cancellation was the AFT Weibull model. Given the structure of the observed time from inception of the contract right-censoring was used in the context of death of the insured, ending of the contract or continuation of the contract in force. Only observations relating to cancellation of the contract were treated as non-censored. For a particular client risk profile we analysed in the paper the effect for example of premium payment frequency on the time to cancellation. Yearly payment of premiums "extends" the time to cancellation by 27% compared with contracts with premiums paid monthly other things being equal. For monthly payment of premiums 10% of clients cancel within approximately 376 days whereas for yearly premium payments it is 476 days. If the client took out a contract with yearly payment of premiums through a bank or insurance broker they would cancel within approximately 343 days which is a 28% shorter time than if they had bought via a tied agent, assuming the other parameters of the contract remain unchanged. Cancellation of insurance contracts has a significant impact on the cash-flows and profitability of the insurer and therefore modelling of cancellations is one of the key aspects of actuarial analyses. A notable challenge in addressing survival analysis lies in the implementation of machine learning methods. Beyond achieving improved predictive accuracy, a key consideration in actuarial analyses is ensuring model interpretability, particularly with respect to quantifying the influence of individual risk factors on the predicted variable.

ACKNOWLEDGMENT

The paper was supported by the grant agency of the Ministry of Education, Science and Research of the Slovak Republic VEGA No. 1/0497/25 *Implementation of innovative approaches in managing and modelling of risk in internal models of insurance companies*, VEGA No. 1/0096/23 *Selected methods of risk management in the implementation of partial internal models for determining the solvency capital requirement*, VEGA No. 1/0124/24 *Slovakia in the context of the pension index and population ageing index – future perspective.*

References

ABADI, A., YAVARI, P., DEHGHANI-ARANI, M., ALAVI-MAJD, H., GHASEMI, E., AMANPOUR, F., BAJDIK, C. (2014). Cox models survival analysis based on breast cancer treatments. *Iranian Journal of Cancer Prevention,* 7(3): 124–129.

AZZONE, M., BARUCCI, E., GIUFFRA, G., MARAZZINA, D. (2021). A machine learning model for lapse prediction in life insurance contracts [online]. *Expert Systems with Applications,* 191. <https://doi.org/10.1016/j.eswa.2021.116261>.

AZIZ, N., RAZAK, S. A. (2019). Survival analysis in insurance attrition [online]. *AIP Conference Proceedings,* 2184(1). <https://doi.org/10.1063/1.5136402>.

BAART, S. J., BOERSMA, E., RIZOPOULOS, D. (2019). Joint models for longitudinal and time-to-event data in a case-cohort design [online]. *Statistics in Medicine,* 38: 2269–2281. <https://doi.org/10.1002/sim.8113>.

BARNWAL, A., CHO, H., HOCKING, T. (2022). Survival regression with accelerated failure time model in XGBoost [online]. *Journal of Computational and Graphical Statistics,* 31(4): 1292–1302. <https://doi.org/10.1080/10618600.2022.2067548>.

COLLETT, D. (2015). *Modelling survival data in medical.* 3rd Ed. Bristol, UK: Taylor and Francis Group, LLC.

ELING, M., KIESENBAUER, D. (2014). What policy features determine life insurance lapse? An analysis of the German market [online]. *Journal of Risk and Insurance*, 81(2): 241–269. <https://doi.org/10.1111/j.1539-6975.2012.01504.x>.

KASARANENI, B. P. (2024). Machine learning techniques for predicting lapse behaviour in life insurance: Advanced models and real-world applications. *Journal of AI-Assisted Scientific Discovery,* 4(1).

KASSAMBARA, A., KOSINSKI, M. (2021). *Survminer: Drawing survival curves using 'ggplot2'* [online]. R package version 0.4.9. <https://CRAN.R-project.org/package=survminer>.

KIESENBAUER, D. (2012). Main determinants of lapse in the German life insurance industry [online]. *North American Actuarial Journal*, 16(1): 52–73. <https://doi.org/10.1080/10920277.2012.10590632>.

KIM, C. (2005). Modeling surrender and lapse rates with economic variables [online]. *North American Actuarial Journal*, 9(4): 56–70. <https://doi.org/10.1080/10920277.2005.10596225>.

KLEIN, D. G., KLEIN, M. (2012). *Survival analysis: a self-learning text*. 3rd Ed. New York: Springer.

KLEIN, J. P., MOESCHBERGER, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*. New York: Springer-Verlag.

KLEIN, J. P. et al. (2014). *Handbook of survival analysis*. USA: Taylor and Francis Group, LLC.

KOMARA, S., ZELINOVÁ, S. (2024). Interactive survival analysis of lung cancer data using Python and Streamlit. In: *Implementácia inovatívnych prístupov modelovania rizík v procese ich riadenia v interných modeloch poisťovní v kontexte s požiadavkami direktívy Solvency II*, Bratislava: EKONÓM.

LI, H. (2017). Survival analysis for a breast cancer data set [online]. *Advances in Breast Cancer Research,* 6: 1–15. <https://doi.org/10.4236/abcr.2017.61001>.

MAJEED, A. F. (2020). Accelerated failure time models: an application in insurance attrition. *The Journal of Risk Management and Insurance,* 24(2).

MILHAUD, X., DUTANG, C. (2018). Lapse tables for lapse risk management in insurance: a competing risk approach [online]. *European Actuarial Journal*, 8(1): 97–126. <https://doi.org/10.1007/s13385-018-0165-7>.

MOORE, D. F. (2016). *Applied survival analysis using R*. Switzerland: Springer International Publishing.

RAMEZANKHANI, A., TOHIDI, M., AZIZI, F., HADAEGH, F. (2017). Application of survival tree analysis for exploration of potential interactions between predictors of incident chronic kidney disease: a 15-year follow-up study [online]. *Journal of Translational Medicine,* 15(1). <https://doi.org/10.1186/s12967-017-1346-x>.

R CORE TEAM (2022). *R: A language and environment for statistical computing* [online]. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org>.

RECK, L., SCHUPP, J., REUß, A. (2022). Identifying the determinants of lapse rates in life insurance: an automated Lasso approach [online]. *European Actuarial Journal,* 13(2): 541–569. <https://doi.org/10.1007/s13385-022-00325-1>.

SAIKIA, R., BARMAN, M. P. (2017). A review on accelerated failure time models. *International Journal of Statistics and Systems*, 12(2): 311–322.

SHENG, J., QIAN, X., RUAN, T. (2018). Analysis of influencing factors on survival time of patients with heart failure [online]. *Open Journal of Statistics,* 8(4): 651–659. <https://doi.org/10.4236/ojs.2018.84042>.

ŠTĚPÁNEK, L., HABARTA, F., MALÁ, I., MAREK, L. (2021). An alternative to Cox's regression for multiple survival curves comparison: a random forest-based approach using covariate structure [online]. In: *International Conference on Computing, Computational Modelling and Applications (ICCMA)*, 130–137. <https://doi.org/10.1109/ICCMA53594.2021.00029>.

ŠTĚPÁNEK, L., HABARTA, F., MALÁ, I., ŠTĚPÁNEK, L., NAKLÁDALOVÁ, M., BORIKOVÁ, A., MAREK, L. (2023). Machine learning at the service of survival analysis: Predictions using time-to-event decomposition and classification applied to a decrease of blood antibodies against COVID-19 [online]. *Mathematics,* 11(4): 819. <https://doi.org/10.3390/math11040819>.

TEPLANOVÁ, P. (2023). Analýza dĺžky poistných kontraktov modelom zrýchleného času. *Ekonomika a informatika,* Ekonomická univerzita v Bratislave, 21(1): 54–70.

TEPLANOVÁ, P., PÁLEŠ, M. (2021). Analýza doby trvania poistných zmlúv využitím analýzy prežitia v jazyku R. *Trendy vo vzdelávaní* študentov študijného *programu Aktuárstvo.* České Budějovice, Vysoká škola evropských a regionálních studií, 109–114.

THERNEAU, T. M. (2023). *A package for survival analysis in R* (version 3.5–7) [online]. <https://CRAN.R-project.org/package=survival>.

YANG, C., DIAO, L., COOK, R. (2021). Survival trees for current status data. *Proceedings of Machine Learning Research,* 146: 83–94.

ZELINOVÁ, S. (2021). Vplyv zvýšenia storna na hodnoty v modeli VFA. *Economics and Informatics*, 19(2): 125–136.

WEI, L. J. (1992). The Accelerated Failure Time Mode: a Useful Alternative to the Cox Regression Model in Survival Analysis [online]. *Statistics in Medicine*, 11: 1871–1879. <https://doi.org/10.1002/sim.4780111409>.

WICKHAM, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.