# MODELLING OF CANCELLATION OF INSURANCE CONTRACTS USING SURVIVAL ANALYSIS

**David Zapletal**

**Abstract**

Several studies have been published showing the link between insurance development and economic growth. The insurance development has been determined by aggregated measures such as insurance density, insurance penetration, insurance premiums, etc. The presented study investigates the link between macroeconomic indicators (such as GDP, unemployment and inflation rate) and cancellation of individual insurance contracts by the clients. The survival analysis is applied to real data from the Czech Republic provided by a commercial insurance company. Given that the macroeconomic indicators evolve over time, it is impossible to use the classical Cox proportional hazards model. Therefore, the extended Cox model with time-dependent variables has to be used. In addition, to examine the impact of macroeconomic indicators, the influence of the gender of an insured person, age at which the person entered into an insurance contract, and region where the insured person lived on the lapse of the insurance contract are studied. It is shown that all considered variables significantly influence the risk of cancellation of the insurance contract. The Cox model enables to quantification of the risk and to estimate the survival probabilities over the studied period.

**Keywords:** critical illness insurance, survival analysis, extended Cox model

**JEL Code:** C14, C24, C41, C51

## Introduction

The link between insurance development and economic growth has been studied by various authors. Beck and Webb (2003) investigated economic, demographic, and institutional determinants of life insurance consumption. Using panel data aggregated at different frequencies for 68 economies in 1961-2000, they found out that economic indicators such as inflation, income per capita, and banking sector development are the most robust predictors of the use of life insurance. Haiss and Sumegi (2008) studied the relationship between insurance and economic growth in 29 European countries from 1992 to 2005. Their findings emphasize the impact of the real interest rate and the level of economic development on the insurance-

growth nexus. Liyan et al. (2010) also investigated this link via a panel dataset of 77 developed and developing economies. They found that insurance development is positively correlated with economic growth. Disaggregated data on real insurance premiums, including life and non-life insurance premiums, were used by Lee (2011) to examine the relationship between insurance markets' activities and economic growth for ten selected OECD countries from 1979 to 2006. They concluded that there is fairly strong evidence favoring the hypothesis of a long-run equilibrium relationship between real GDP and insurance markets' activities after allowing for the heterogeneous country effect. Their long-run panel regression parameter results also indicated a significant positive relationship between real GDP and the activities within the insurance market. By implementing the dynamic panel-based error correction model, they determined that insurance markets' development and economic growth present both the long-run and short-run bidirectional causalities. The causal relationships between insurance market development, financial development, and economic growth in 34 OECD countries for the period 1988–2012 were found by Pradhan et al. (2015). The link between insurance market penetration and per capita economic growth in 19 Eurozone countries from 1980 to 2014 was studied by Dash et al. (2017). Their empirical results perceive both unidirectional and bidirectional causality between insurance market penetration and per capita economic growth.

Most of the above studies assessed the insurance development by aggregated measures such as insurance density, insurance penetration, insurance premiums, etc., and they used methods for the analysis of multivariate time series. The presented paper investigates the link between macroeconomic indicators (in this case, unemployment and inflation rate) and cancellation of individual insurance contracts by the clients using survival analysis. The analysis was applied to real data from the Czech Republic provided by the commercial insurance company.

Survival analysis is an area of statistics that estimates the amount of time that is likely to elapse before a certain event occurs. Models for this type of analysis have often been applied to medical data, and many scientific books and research articles have reported on the results. The use of survival analysis in the insurance industry for contract failures is also quite common. For instance, the analysis of customer survival time in an insurance company after a policy cancellation was introduced by Guillen et al. (2003). The Cox proportional hazards model was used by Ho and Su (2006) to investigate China's residential mortgage life insurance prepayment risk behavior. The competing risks approach was used by Mihaud and Dutang (2018). They modeled the duration of a life insurance contract through the subdistribution hazard model.

# 1 Methodology

The most widely used regression model in survival analysis is the Cox proportional hazards model, which provides a suitable method for accommodating covariate information. The model, with its unspecified baseline hazard function, was proposed by Cox (1972), and the notion of partial likelihood was introduced. The Cox model of the hazard at time *t* is given by the equation

$$h(t) = \exp(\boldsymbol{\beta}\mathbf{X})\, h_0(t), \tag{1}$$

where $h_0(t)$ is the baseline hazard function of the unspecified form. An essential feature of Eq. (1), which concerns the proportional hazards (PH) assumption[1], is that the baseline hazard is a function of time but does not involve the covariates X′s, whereas the exponential expression involves the X′s but does not involve time. The X′s here are called time-independent.

It is possible, nevertheless, to consider X′s that do involve time (or change over time); such X′s are called time-dependent variables. If time-dependent variables are considered, the Cox model form may still be used, but such model no longer satisfies the PH assumption and is called the extended Cox model (Kleinbaum and Klein, 2012). Of course, in addition to time-dependent variables, time-independent variables can also appear in the extended Cox model. However, such variables should meet the PH assumption. The extended Cox model (containing both time-independent and time-dependent variables) of the hazard at time *t* has the form

$$h(t) = \exp[\boldsymbol{\beta}\mathbf{X} + \boldsymbol{\delta}\mathbf{X}(t)]\, h_0(t), \tag{2}$$

where $h_0(t)$ is the baseline hazard function and $\mathbf{X} = (X_1, \dots, X_k)$ denotes the vector of time-independent variables and $\mathbf{X}(t) = (X_1(t), \dots, X_p(t))$ the vector of time-dependent variables.

# 2 Data

A commercial insurance company in the Czech Republic provided data on critical illness insurance. This type of insurance is offered as a supplement to life insurance if the policyholder is an adult under the age of 65 or as a separate insurance contract for persons under 18 years of age. The policyholders under 18 were not included in the analysed dataset. The insurance covers the risk of 31 critical illnesses such as cancer, heart attack, and stroke. If the client is diagnosed with any of these diseases, the client will receive the agreed sum insured. The insurance indemnity is paid only once, and after its payment, the insurance contract expires.

[1] The PH assumption underlying the Cox PH model means that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant (hazard ratio – HR) is independent over time.

The analysed dataset in the study contained clients who entered into insurance contracts during 2010 (i.e. from January 2010 to December 2010). The follow-up period of these contracts then lasted until April 2017. Therefore, the date of the start point and eventual endpoint of each policy was known. If the endpoint was reached on or before April 30, 2017, its cause was known. The endpoint could be the date that an insured event occurred or the date that the policy was terminated by the client or insurance company. In this study, attention was focused on situations where the insurance was terminated unexpectedly (for the insurance company), which brought a financial loss to the insurance company. Such situations included termination of the insurance by the client, termination of the insurance for non-payment of insurance premiums, and the occurrence of an insured event. All these cases were considered as *events* in the present analysis. Other causes of termination of the insurance contract (for example, expiration of the contract) or policy in force were deemed to be *censored* cases.

In order to analyse the time to the endpoint (i.e., to the lapse of the insurance contract), the duration of each policy (in months) was calculated. In addition to the policy duration (including the date of the beginning and eventual end of the insurance contract), we also had information about the gender and age of the policyholder and the region in which the policyholder lived. This information was then supplemented by the relevant monthly values of the inflation rate and unemployment rate in the Czech Republic. This data was obtained from ARAD data series system (2021), forming part of the Czech National Bank's information service. Since macroeconomic factors usually have delayed effect, time lags of 6 months were introduced for inflation rate and unemployment rate. Therefore, e.g., the effect of inflation is examined in January 2015 on the possible termination of the insurance contract in July 2015.

The analysed dataset contains information of clients who took out insurance during 2010, and this insurance lasted for more than one year. The limitation is made based on our study of the effect of macroeconomic factors on the cancellations. It can be assumed that the main reason for cancelled insurance contracts during the first year or exactly after one year was most likely not the effect of macroeconomic factors. The final dataset contained data for 11,030 persons, and the number of events in the monitored period was 5,470.

# 3 Model

Three time-independent variables (gender of the insured person, age at which the person entered into the insurance contract, and region where the insured person lived) and two time-dependent variables (inflation rate and unemployment rate) were included in the model. The description

of individual variables is given in Tab. 1. Because of the categorical (factor) time-independent variables, each variable is represented by $q$-1 dummy variables, where $q$ means the number of categories of corresponding explanatory variables (see Tab. 1). The variable Gender included two categories. The Age variable was made up of three categories: 18 to 28 years, 29 to 45 years, and over 45 years. Fourteen regions of the Czech Republic were divided into three categories of variable GrRegions. These categories were formed based on the values of median survival time, where the regions with the highest median survival time form the 1st group and those with the lowest median survival time form the 3rd group. The categories of corresponding variables with the lowest median survival time were determined as reference categories; they were male for variable Gender, over 45 years for variable Age, and the 1st group for variable GrRegions. The reference categories are indicated in bold in the third column of Tab. 1.

In addition, the full model also included the interaction of macroeconomic factors (time-dependent variables) and individual time-independent variables. Therefore, extended Cox (full) model of the hazard at time $t$ is, in this case, given by the equation

$$\begin{aligned} h(t) = \exp[&\beta_1 Gender(Male) + \beta_2 Age(18-28) + \beta_3 Age(29-45) + \beta_4 GrRegions(2nd) \\ &+ \beta_5 GrRegions(3rd) + \delta_1 CPI(t) + \delta_2 Unemp(t) + \gamma_1 Gender(Male){:}\, CPI(t) \\ &+ \gamma_2 Age(18-28){:}\, CPI(t) + \gamma_3 Age(29-45){:}\, CPI(t) + \gamma_4 GrRegions(2nd){:}\, CPI(t) \\ &+ \gamma_5 GrRegions(3rd){:}\, CPI(t) + \gamma_6 Gender(Male){:}\, Unemp(t) \\ &+ \gamma_7 Age(18-28){:}\, Unemp(t) + \gamma_8 Age(29-45){:}\, Unemp(t) \\ &+ \gamma_9 GrRegions(2nd){:}\, Unemp(t) + \gamma_{10} GrRegions(3rd){:}\, Unemp(t)] \cdot h_0(t). \end{aligned} \quad (3)$$

For the variables used in extended Cox model as time-independent, a graphical check, through Kaplan-Meier survival curves, for the PH assumption was performed and showed that this assumption was not violated.

The inclusion of time-dependent variables in the model requires special attention. The follow-up period of each policy had to be divided into time intervals according to ordered survival (event) times. Since at least one event occurs in each month of the follow-up period, this period was divided into monthly intervals to which the corresponding (delayed) values of macroeconomic factors were assigned. For example, the duration of the insurance cancelled by the client after three years was divided into 36 intervals. Afterwards, the first 35 intervals were considered as censored cases, and the last interval ended with the observed event. The tmerge function (Therneau et al., 2020), which is implemented in the R-software as a part of the survival package, was used for data preparation in this way. With this transformation, the size of the data set increased from the original 11,030 cases to 521,134 observations. All other calculations were also performed using the survival package (Therneau, 2020) in R.

**Tab. 1: Explanatory variables of the model**

| Variable name | Description | Categories | Time dependence |
|---|---|---|---|
| Gender | | **Female**, Male | Time-independent |
| Age | Age of the client when setting up the insurance | 18 – 28; 29 – 45; **over 45** | Time-independent |
| GrRegions | Groups of regions based on values of median survival time | **1st:** OLK, PAK, PHA, VYS, ZLK<br>2nd: LBK, STC, HKK, MSK, JHM<br>3rd: KVK, ULK, PLK, JHC | Time-independent |
| CPI | Inflation rate in %<br>– lag 6mth | continuous variable | Time-dependent |
| Unemp | Unemployment rate in %<br>– lag 6mth | continuous variable | Time-dependent |

Source: own

# 4 Results

As mentioned above, the full model given by Eq. (3) was first considered, but then a subset of explanatory variables was selected by reducing the full model in a stepwise fashion based on the Akaike information criterion (AIC). Therefore, the final model contains all five main effects (three time-indep. and two time-dep.) and four interaction variables and has the form

$$\begin{aligned} h(t) = \exp[&\beta_1 Gender(Male) + \beta_2 Age(18-28) + \beta_3 Age(29-45) + \beta_4 GrRegions(2nd) \\ &+ \beta_5 GrRegions(3rd) + \delta_1 CPI(t) + \delta_2 Unemp(t) + \gamma_2 Age(18-28){:}\, CPI(t) \\ &+ \gamma_3 Age(29-45){:}\, CPI(t) + \gamma_4 GrRegions(2nd){:}\, CPI(t) + \gamma_5 GrRegions(3rd){:}\, CPI(t) \\ &+ \gamma_6 Gender(Male){:}\, Unemp(t) + \gamma_7 Age(18-28){:}\, Unemp(t) \\ &+ \gamma_8 Age(29-45){:}\, Unemp(t)] \cdot h_0(t). \end{aligned} \quad (4)$$

Estimations of coefficients of the extended Cox model, given by the Eq. (4), with their robust standard errors[2] and statistical significance (p-values), are shown in Tab. 2. We can see that both considered macroeconomic indicators significantly (for $\alpha = 0.05$) influence the risk of cancellation of the insurance contract. In addition, due to the statistical significance of the interaction effects, we can say that the influence of macroeconomic factors is different for individual categories compared to the reference categories. However, to quantify this effect some additional calculations need to be done.

[2] These robust standard errors are designed to account for the non-independence of observations from the same subject. Because due to inclusions of the time-dependent variables there are multiple observations (clusters) from the same subject.

**Tab. 2:** Estimation results of the extended Cox model (rounded to four decimal places).

| Variable name | Coeff. | Rob. stand. error | p-value | Variable name (interactions) | Coeff. | Rob. Stand. Error | p-value |
|---|---|---|---|---|---|---|---|
| Gender(Male) | -0.4848 | 0.2037 | 0.0173 | Age(18-28):CPI | -0.1932 | 0.0481 | 0.0000 |
| Age(18-28) | 0.3021 | 0.3606 | 0.4022 | Age(29-45):CPI | -0.0949 | 0.0464 | 0.0407 |
| Age(29-45) | -0.5616 | 0.3513 | 0.1099 | GrRegions(2nd):CPI | -0.0154 | 0.0271 | 0.5709 |
| GrRegions(2nd) | 0.1917 | 0.0549 | 0.0005 | GrRegions(3rd):CPI | 0.1503 | 0.0329 | 0.0000 |
| GrRegions(3rd) | 0.0691 | 0.0718 | 0.3361 | Gender(Male):Unemp | 0.0759 | 0.0311 | 0.0147 |
| CPI | 0.1103 | 0.0522 | 0.0344 | Age(18-28):Unemp | 0.0729 | 0.0623 | 0.2414 |
| Unemp | -0.1606 | 0.0809 | 0.0471 | Age(29-45):Unemp | 0.1415 | 0.0606 | 0.0195 |

Source: own

The so-called hazard ratio (HR) is often used to quantify and interpret the results of survival models. Because inflation and unemployment were included in the model as time-dependent variables, the HR related to these variables is also time-dependent. If we are interested in, e.g., how unemployment affects the risk of cancellation of insurance for males compared to females (ref.), we need to calculate the point estimate of HR, which is given by

$$\widehat{HR}(t) = \exp\left[\hat{\beta}_1 \cdot 1 + \hat{\gamma}_6 \cdot 1 \cdot Unemp(t)\right]. \tag{5}$$

Here $\hat{\beta}_1$ and $\hat{\gamma}_6$ are the estimated coefficients from Eq. (4) (see Tab. 2 for specific values) and $Unemp(t)$ means the (6 months delayed) unemployment rate. For example, if the unemployment was low (6 months ago), say 3%, the $\widehat{HR}(t) = 0.7732$. Of course, to assess the statistical significance of the estimate, it is necessary to construct a confidence interval (or test significance) of HR. Generally, the (100-$\alpha$)% confidence interval for the true HR is given by

$$\exp\left(\widehat{HR}(t) \pm u_{1-\alpha/2} \cdot \sqrt{Var\left[\widehat{HR}(t)\right]}\right), \tag{6}$$

where $u_{1-\alpha/2}$ is a quantile of standard normal distribution and $Var\left[\widehat{HR}(t)\right]$ is the variance of the estimate of HR. Because we want to estimate the variance of the interaction variable, we need to know both variances of the estimated coefficients and their covariance. In the case of the estimate of HR given by Eq. (5) the variance is calculated in the form

$$Var\left[\widehat{HR}(t)\right] = Var\left(\hat{\beta}_1\right) + [Unemp(t)]^2 \cdot Var(\hat{\gamma}_6) + 2 \cdot Unemp(t) \cdot cov\left(\hat{\beta}_1, \hat{\gamma}_6\right). \tag{7}$$

For $Unemp(t)$ equal to 3% the $Var\left[\widehat{HR}(t)\right] = 0.1120$, where $Var\left(\hat{\beta}_1\right)$, $Var(\hat{\gamma}_6)$ and $cov\left(\hat{\beta}_1, \hat{\gamma}_6\right)$ were obtained from the variance-covariance matrix, which is not presented here. Therefore, the corresponding 95% confidence interval for true HR is $(0.6208; 0.9630)$. Because the confidence interval does not include value one, we can say that for the unemployment rate of 3% the risk

of cancellation of insurance policy is significantly lower for males compared to females. Because $\hat{\gamma}_6$ in Eq. (5) has a positive value ($\hat{\gamma}_6 = 0.0759$), for the increasing unemployment rate is $\widehat{HR}(t)$ also increasing. Therefore, for the unemployment rate approximately from 5% to 8%, there is no significant difference in risk of cancellation between males and females. But for unemployment over 8%, the risk of lapse of the policy is higher for males. For example, with $Unemp(t) = 10\%$ is $\widehat{HR}(t) = 1.3150$ with the confidence interval $(1.0549; 1.6391)$. Of course, similar calculations can be done for other interaction effects.

We can investigate how the hazard ratio is influenced by the unemployment rate for age groups 18-28 and 29-45 years compared to the reference age group (over 45 years). Because both $\hat{\gamma}_7$ and $\hat{\gamma}_8$ is positive, the estimate of HR is increasing with increasing unemployment rate. In the age group 18–28 years, HR is significantly higher than one for unemployment around 3%, while in the age group 29–45 years, it is for the value of unemployment around 6%. On the other hand, if the comparison between age groups 18-28 and 29-45 (reference) is made[3], the HR is decreasing with increasing unemployment, and it is significantly higher than one for low unemployment rates and insignificant for unemployment around 9%.

As regards inflation, the impact on the risk of cancelling insurance varies mainly regionally. While in the 3rd region group compared to the reference 1st one, the risk is increasing with increasing inflation, in the 2nd group of regions, this risk is slightly decreasing. For the 2nd group in case of deflation ($\mathrm{CPI} = -1\%$) the $\widehat{HR}(t)$ is equal $1.2301$ with confidence interval $(1.0540; 1.4356)$ and for higher inflation rate (in developed countries), say 5%, the $\widehat{HR}(t)$ is insignificant. On the other hand, in the 3rd group, in case of deflation ($\mathrm{CPI} = -1\%$) the $\widehat{HR}(t)$ is insignificant and for the inflation rate of 5% the $\widehat{HR}(t) = 2.2714$ with confidence interval $(1.8247; 2.8275)$. If the 3rd and 2nd (reference) groups are compared to each other, the HR increases with rising inflation from significantly lower than 1 ($\widehat{HR}(t) = 0.7495$) for deflation ($\mathrm{CPI} = -1\%$) to significantly higher than 1 ($\widehat{HR}(t) = 2.0245$) in case of an inflation rate of 5%.

In the case of age groups, the increase of inflation should decrease the risk of cancellation for both 18-28 and 29-45 years compared to the reference group (over 45 years). But for realistically possible values of inflation in the Czech Republic (say, from -1% to 5%), most of the estimates of HR are insignificant. On the other hand, for comparison between age groups 18-28 and 29-45 (reference), the HR decreases with increasing inflation. It is significantly higher than one for deflation (and low inflation) and insignificant for inflation around 2%.

[3] The model was re-estimated with the reference age group 29-45 years. Results does not presented here.

## Conclusion

The link between macroeconomic indicators (unemployment and inflation rate) and the cancellation of individual insurance contracts by the clients was studied. The extended Cox model with both time-dependent and time-independent variables was applied to real data from the Czech Republic provided by the commercial insurance company. Especially in connection with the statistical significance of interaction effects, several interesting conclusions can be drawn.

If the unemployment rate is low, say around 3%, the risk of cancelling the insurance is approximately 1.3 times higher for women than men. On the other hand, with the rising unemployment, this risk is balanced, and from around 8% onwards, this risk is already greater for men and is growing. This might be due to the fact that the worse economic situation is leading men to increase their income by cancelling the insurance. However, it should be noted that with the usual unemployment rates in the Czech Republic, the difference between men and women is not so significant. The unemployment would have to rise to 16% to double this risk.

The result of the unemployment and age group interactions shows that the most sensitive age group for changes in the unemployment rate is the group 29-45 years. If the unemployment rate is low, the risk of cancelling the insurance for this group is comparable to the age group over 45 years. On the contrary, with a higher unemployment rate, the risk of age group 29-45 years is similar to the highest risk group 18-28 years.

Regarding the inflation rate, in case of deflation or zero inflation, the risk of cancellation of insurance varies only slightly between regions but significantly between age groups 18-28 and 29-45 years. But rising inflation causes an increase in risk for the 3rd region group (with the lowest median survival time) compared to both 1st and 2nd group of regions, but decrease to insignificant differences in risk for the age groups.

## References

ARAD data series system of the Czech National bank (2021) https://www.cnb.cz/cnb/STAT.ARADY_PKG.DATOVE_ZDROJE

Beck, T, & Webb, I. (2003). Economic, demographic, and institutional determinants of life insurance consumption across countries. *World Bank Economic Review*, 17(1), 51-88.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of Royal Statistical Society B*, 34(2), 187-220.

Dash, S., Pradhan, R. P., Maradana, R. P., Gaurav, K., Zaki, D. B., & Jayakumar, M. (2017). Insurance market penetration and economic growth in Eurozone countries: time series evidence on causality. *Future Business Journal*, 4, 50-67.

Guillen, M., Nielsen, J. P., Parner, J. T., & Perez-Marin, A. M. (2003). The analysis of customer survival time in the insurance company after a policy cancellation. *Insurance Mathematics & Economics*, 33, 434-434.

Haiss, P., & Sumegi K. (2008). The relationship between insurance and economic growth in Europe: a theoretical and empirical analysis. *Empirica*, 35, 405-431.

Ho, K. H., & Su, H. Y. (2006). Structural prepayment risk behavior of the underlying mortgages for residential mortgage life insurance in a developing market. *Journal of Housing Economics*, 15, 257-278.

Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis. A self-learning text.3rd ed.* Springer.

Lee, Ch. Ch. (2011). Does insurance matter for growth: empirical evidence from OECD countries. *The B.E. Journal of Macroeconomics*, 11(1), 1-28.

Liyan, H, Donghui, L, Fariborz, M, & Yanhui, T. (2010). Insurance Development and Economic Growth. *The Geneva Papers on Risk and Insurance*, 35, 183-199.

Mihaud, X., & Dutang, Ch. (2018). Lapse tables for lapse risk management in insurance: a competing risks approach. *European Actuarial Journal*, 8, 97-126.

Pradhan, R. P., Arvin, M. B., & Norman, N. R. (2015). Insurance development and the finance-growth nexus: Evidence from 34 OECD countries. *Journal of Multinational Financial Management*, 31, 1-22.

Therneau, T. (2020, March 7). *A Package for Survival Analysis in R. R package version 3.2-10*. https://cran.r-project.org/web/packages/survival/index.html.

Therneau, T., Crowson, C., & Atkinson, E. (2020, September 25). *Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model*. https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf

## Contact

David Zapletal
Faculty of Economics and Administration, University of Pardubice
Studentská 95, 53210 Pardubice
david.zapletal@upce.cz