RUHUNA JOURNAL OF SCIENCE
Vol 11 (2): 118-130, December 2020
eISSN: 2536-8400
DOI: http://doi.org/10.4038/rjs.v11i2.91

Modelling non-life insurance in Sri Lanka using Cox Hazard Model and classification of risky customers

W.A.R. De Mel and W.A.P.A. Chathurangani

Department of Mathematics, University of Ruhuna, Wallamadama, Matara 81000, Sri Lanka

Correspondence: piumiayodyachathurangani@gmail.com; https://orcid.org/0000-0001-8296-9218

Received: 20th June 2020, Revised: 25th December 2020, Accepted: 30th December 2020

Abstract. Some of the major factors that help the decision-making process of an insurance company include Time of the first claim (TFC), claim Size and the frequency of claims. However, in most situations researchers focus mainly on the second and third factors mentioned above. We hypothesize the importance of the TFC of an insurance contract in the decision-making process. Empirical evidence of motor vehicle insurance data in Sri Lanka suggests that nine covariates are responsible for the claim sizes. In the current study, our main objective is to find the key factors of those nine that are responsible for the TFC of the insurance contract. This study is based on the claim data in the whole year of 2016 of non-life insurance policies of a particular insurance company in Sri Lanka. Considering the TFC as right-censored data, selected nonparametric methods, i.e., Kaplan-Meier, Nelson-Aalen estimators, and Cox Proportional Hazard Model are used to analyze the data. We identified the five most influential covariates namely, vehicle type, log of Premium Value and that of Assured Sum, the lease type and the age range via fitting the Cox Model to TFC data. After a thorough residual analysis, the Logistic regression model has been used to identify the important covariates to classify future customers as risky or not.

Key words: Classification, Cox Proportional Hazard model, Kaplan-Meier, Right-censored

1 Introduction

In medical follow-up studies, life testing, insurance, and other fields, it is impossible to observe the lifetime or the time of the first insurance claim of all subjects in the study. The reason is time itself. In most cases, it is highly likely that all the events have not been observed by the time one wants to analyze these lifetimes. For example, in a non-life insurance, not every insurance contract has a claim during its contract period.

The individuals in the study who have no claims by the end of the study (or contract) period are labeled as right-censored. The only information the researcher has is the time between the contract initiation and the end of the study time, which is naturally less than the time to the first claim. The simplest kind of censoring is single censoring

which occurs when all subjects are censored at the same time. There are two types of censoring, namely Type I and Type II censoring. In Type I, the censoring time is predetermined whereas Type II occurs when a predetermined number of failures are observed, and the remaining subjects are then right censored. In many studies, subjects are not censored at the same time. This is called random censoring. These types of data are commonly referred to as survival data. The analysis of such data is important in many fields including reliability, engineering, biology, insurance, and medicine (De Mel 2014).

Under the random censoring model, we assume that $X_1, X_2, \dots, X_n$ are independent nonnegative random variables with an absolutely continuous distribution function $F(t) = P(X \leq t)$. The censoring variables $Y_1, Y_2, \dots, Y_n$ are also independent nonnegative random variables with an absolutely continuous distribution function $G(y) = P(Y \leq y)$. We further assume that both random variables X and Y are independent from each other. In this model, the observable random variables are $Z_i = \min(X_i, Y_i)$ and $\delta_i = I(X_i \leq Y_i)$, where δ_i indicates whether Z_i is an uncensored observation or not. The Y_i's right censor the X_i's.

In survival analysis, a variety of parametric and nonparametric significant tests can be used to identify the observed differences among the empirical survival curves. The most commonly used nonparametric test is based on logrank statistic (Oulidi *et al*. 2010). In survival literature, the survival function is usually estimated from the observed data by using the Kaplan-Meier estimator (Kaplan *et al*. 1958). Nelson-Aalen proposed a nonparametric estimator (Nelson 1972) for the cumulative hazard rate function $\Lambda(t)$.

In non-life insurance studies, one of the main interests is to investigate the association between the claim sizes or claim times of insurance contracts and related covariates. These covariates are sometimes referred to as risk factors. Examples of commonly encountered risk factors include age, sex and the health condition of the contract holder, vehicle type, premium value, hiring status, assured sum, lease type, and the brand name of a vehicle. Identifying and measuring this association helps insurance companies to understand how these factors are associated with the occurrence or nonoccurrence of an insurance claim. This in turn helps insurance companies to calculate the premium of a contract according to the customer's risk.

Since the survival data is not normally distributed and has partial information, the standard statistical techniques like multiple linear regressions cannot directly be applied to analyze such data. Cox proposed a semi-parametric multiple regression models for survival data called Cox's Proportional Hazard Model (Cox 1972). This model can be used to identify the important factors in the study and to compare the hazard rate functions among different groups. Furthermore, contrast to parametric models, this method makes no assumptions about the baseline hazard function (Cox 1972).

In the final part of this research, we use the logistic regression model to identify a future customer as a risky or not. This is a classification problem with binary response data variable having categories, 1 for a risky customer and 0 if not risky. In machine learning literature the most commonly used model for this purpose is the Logistic Regression model, a type of Multiple Linear Regression model. In insurance, we

collect the data in only the insurance policy period. Therefore, we can classify each customer as a risky customer or not without any difficulty.

2 Materials and Methods

In this section, we introduce some basic concepts and their definitions in survival analysis

2.1 Survival Function

Let T be an arbitrary continuous nonnegative random variable with distribution function $F(t) = P(T \leq t)$ and the density function $f(t) = \frac{dF(t)}{dt}$. Survival function of T is defined as

$$S(t) = \Pr(T > t) = 1 - F(t). \qquad (1)$$

This measures the probability that a subject survives beyond some specific time t. The hazard rate function or instantaneous event rate is usually denoted by $h(t)$ and it is the probability that an individual who is under observation has an event at time t. Define by

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)}. \qquad (2)$$

The function $\Lambda(t) = \int_0^t h(u)du$ is the cumulative hazard function of T and $S(t) = \exp\{-\Lambda(t)\}$.

2.2 Kaplan-Meier Survival Estimator

The Kaplan-Meier (KM) (Kaplan et. al, 1958) is a nonparametric estimator of a survival function $S(t)$ given by

$$\hat{S}(t) = \prod_{k:T_k \leq t} \exp\left(-\frac{d_k}{n_k}\right) = \prod_{k:T_k \leq t} \left(1 - \frac{d_k}{n_k}\right), \qquad (3)$$

where n_k is the number of subjects at risk (alive) just before T_k and d_k is the number of failures at time T_k. The times $T_1 < T_2 < \cdots < T_L$ denote the L distinct ordered observed failure times in the study with n subjects. The Nelson-Aalen estimator (Nelson, 1972) for Λ(t) is given by $\hat{\Lambda}(t) = \prod_{k:T_k \leq t} \frac{d_k}{n_k}$. With tied failure data, the above formulas need to be modified.

2.3 Cox Proportional Hazards Model

The Cox Proportional Hazards Model (Cox 1972) is a semi-parametric model because it does not assume any conditions on $h_0(t)$.

The Cox's model (Cox 1972) is given by

$$h(t|\,X) = h_0(t)\exp\left(\sum_{i=1}^{p}\beta_i X_i\right), \qquad (4)$$

where $h(t|\,X)$ which is usually written as just $h(t)$ is the hazard rate function at time t, $h_0(t)$ the baseline hazard rate function, and $X = (X_1, X_2, \ldots, X_p)$ the covariate vector of size p. This model can be used to analyze survival data by regression model such as multiple linear regression and generalized linear models because equation (4) can be reduced to $\log\left(\frac{h(t|\,X)}{h_0(t)}\right) = \sum_{i=1}^{p}\beta_i X_i$. The dependent variable in the Cox's model is the hazard rate function, $h(t)$. Therefore, one can use this model to compare the survival curves in different groups by taking into account other related covariates.

We assume that $h_0(t)$ is unknown and common to all subjects in the study. Most of the time, the coefficients $\beta_1, \beta_2, \ldots, \beta_p$ are estimated by using the partial likelihood method (Cox, 1972). The partial likelihood function is given by the following equation.

$$PL(\beta) = \prod_{i=1}^{n}\prod_{t\geq 0}\left[\frac{Y_i(t)e^{X_i(t)\beta}}{\sum_{i=1}^{n}Y_i(t)e^{X_i(t)\beta}}\right]^{dN_i(t)} \qquad (5)$$

where $N_i(t) = I(Z_i \leq t, \delta_i = 1)$ and $Y_i(t) = I(X_i \geq t)$ are the event process and at-risk process respectively for i^{th} subject (Fleming et al., 1991). In most of the times, one cannot find the exact solutions to equation (5) but can obtain the numerical solutions for the coefficient vector β using numerical method like Newton-Raphson.

2.4 Logistic Regression Model

To model a binary response variable Y, one can use the logistic multiple regression model. This is given as

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \qquad (6)$$

where $X = (X_1, X_2, \ldots, X_p)$ is vector of covariate, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ vector of regression coefficients, and $p(X) = P(Y = 1\,/X)$. With a little algebra, one can rewrite the equation (6) as

$$p(X) = \frac{\exp\left(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p\right)}{1+\exp\left(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p\right)} \qquad (7)$$

In logistic regression, parameters are estimated by using maximum likelihood method. Future observations can be classified into two groups by using $p(X)$ with some probability threshold, for example $p(X) = 0.5$.

3 Results and Discussion

In this study, we consider a dataset consisting of motor vehicle insurance contracts from a leading insurance company in Sri Lanka. Customers were enrolled for a one-year period from January 01, 2016 to December 31, 2016 therefore; the end of the study period was December 31, 2016. Our dependent variable is the time of the first claim of the contract and we also consider nine covariates, namely, vehicle type, premium value, hiring status, gender, age, assured sum, lease type, brand name of a vehicle, and class type of a vehicle.

Our insurance dataset contains more than 300,000 non-life insurance (motor vehicle insurance) contracts. After removing contracts with missing values we use simple random sampling technique to obtain a random sample with 599 contracts. Any contract with no claims in the study period is treated as right censored data. In Figure 1, we depict a part of our dataset to give some idea about the survival data. Study began on January 01, 2016 and was terminated on December 31, 2016. Insurance contracts can initiate at any time during the contract year. In Figure 1, the horizontal lines depict the claim time of the first claim of an insurance contract. The blue dots on these lines represent the actual time of the first claim and this is known as the calendar time of the first claim.

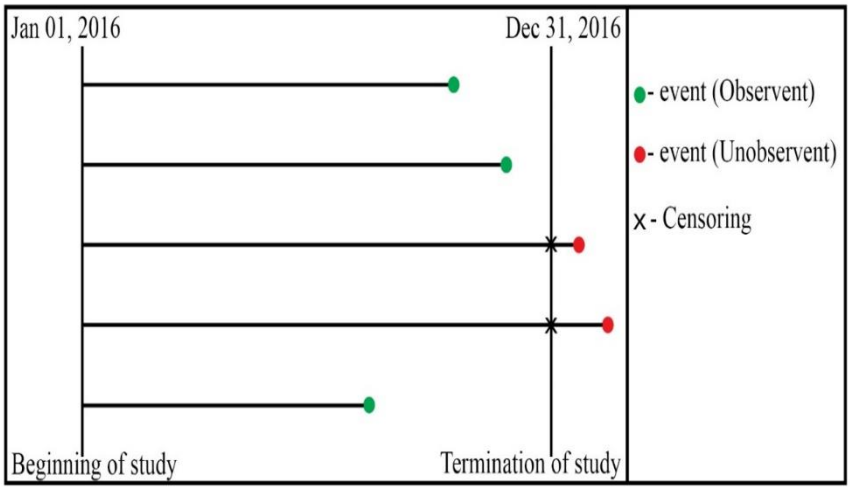

Fig 1: Calendar times of subjects in the study.

If we cannot observe the first claim during the year, these times go beyond the termination time of the study and they are called right censored data. For our dataset, we first compute the Kaplan-Meier survival estimate by using survival package in R software. It is depicted along with the 95% confidence band in Figure 2.

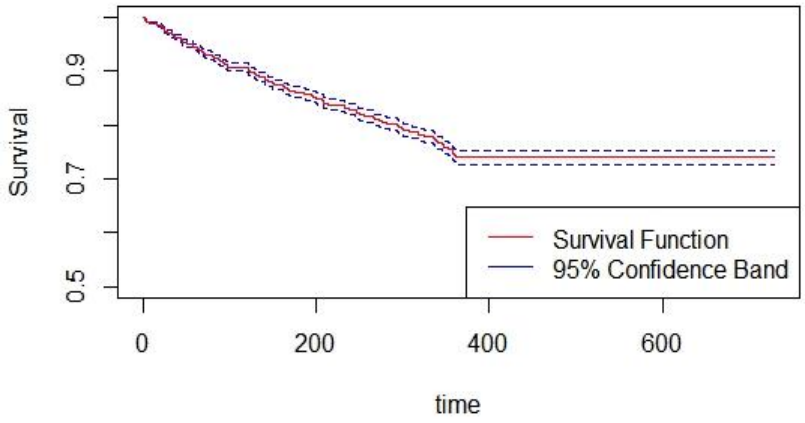

Fig 2: **Kaplan-Meier estimate for claim times for motor vehicle insurance claim data**

Table 1 displays the summary statistics obtained using Logrank test to compare significant differences among the survival curves for categories of categorical variables in the study. All these factors are significant at the 5% significance threshold.

Table 1: Summary statistics for categorical variables based on logrank.

Variable Name	P-value
Age range	0.0421
Lease type	0.00441
Vehicle type	3.69e-11

Figure 3 displays the estimated survival curves for two age ranges. This clearly displays a significant difference between two estimated curves for age range categories and verifies the Logrank result. Here, the two groups for age range are above and below 35 years. But one can use more than two groups. It is clear from the Figure 3 that the youngers are riskier than the elders. Identifying theses risky groups helps insurance companies to calculate the premiums for their insurance contracts.

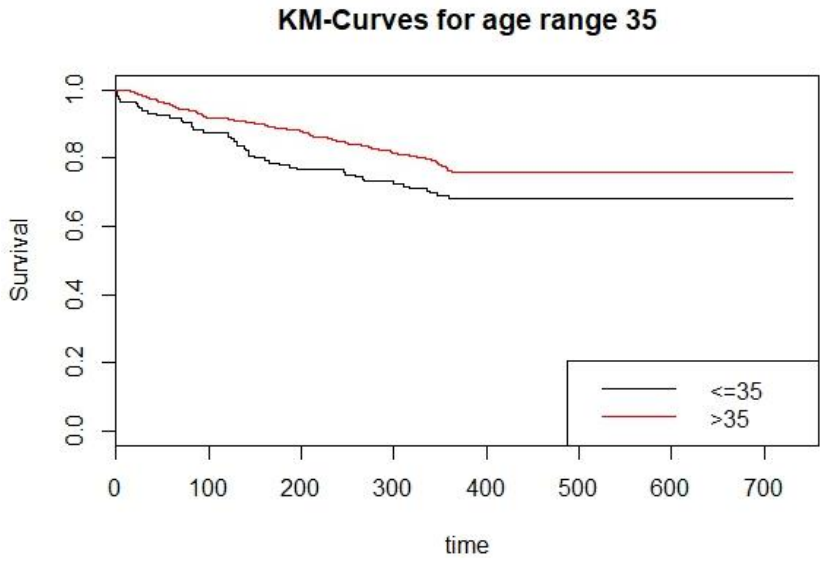

Fig 3: Estimated survival functions by age range

We next fit the Cox Proportional Hazard Model (Cox 1972) including all above mentioned nine covariates. Results from R software are displayed below in Table 2. According to the output, only five variables, namely vehicle type, premium value, assured sum, lease type, and age range are significant at the 5% confidence threshold. Gender is not significant.

After including other six predictors in the model only age range is significant but not gender at the significance level 0.05. We can remove the gender from the Cox Proportional Hazard Model (Cox 1972) since the p-value is larger than the significance level 0.05 but p-value for age range is less that above significance level.

Table 2: Parameter estimates for Cox Proportional Hazard Model with all the variables.

Variable name	β (estimated coefficients)	SE (est. coef.)	Wald test	p-value
Vehicle type	- 9.515e-02	3.341e-02	- 2.848	0.004395
Premium value	3.731e-05	7.627e-06	4.892	1.00e-06
Hiring Status	- 4.213e-02	2.342e-01	- 0.180	0.857234
Assured-sum	- 5.978e-07	1.646e-07	- 3.632	0.000281
Lease type	6.972e-01	1.782e-01	3.912	9.16e-05
Gender	7.823e-02	1.965e-01	0.398	0.690582
Age range	- 5.256e-01	1.813e-01	- 2.900	0.003736
Brand name	3.291e-03	7.609e-03	0.433	0.665334
Class type of vehicle	- 5.872e-03	6.672e-02	- 0.088	0.929873

After removing insignificant predictors, we fit the Cox model again with predictors, namely vehicle type, premium value, assured sum, lease type, and age range. The results from R software for the reduced Cox Proportional Hazard Model are displayed below in Table 3.

Table 3: Parameter estimates for the reduced Cox Proportional Hazard Model.

Variable name	β (estimated coefficients)	SE (est. coef.)	Wald	p-value
Vehicle type	-9.91e-02	2.16e-02	-4.59	4.5e-06
Premium value	3.78e-05	7.49e-06	5.05	4.5e-07
Assured_sum	-6.02e-07	1.61e-07	-3.75	0.00018
Lease type	6.82e-01	1.75e-01	3.89	9.9e-05
Age range	-5.08e-01	1.79e-01	-2.84	0.00447

According to the results in Table 3, all variables in the reduced model are significant at the 5% confidence threshold. Therefore, we could treat this reduced Cox Proportional Hazard Model with variables, namely, vehicle type, premium value, assured sum, lease type, an age range as our final model for the motor vehicle insurance data set. Before using it for prediction purposes it is necessary to carry out the model validation. In model validation, we check the proportional hazard assumption, influential observations assumption and nonlinearity assumption.

3.1 Model validation

Since Cox proportional hazard model fits under several assumptions it is better to check whether the fitted Cox model adequately explains the data. In order to check the three main assumptions namely, violation of the assumption of proportional hazards, influential data and, nonlinearity in the relationship between the log hazard and the covariate, the residuals method is used. Schoenfeld and Martingale residuals (Fleming *et al*. 1991) are the most frequently used residuals to check the assumptions of the Cox model.

Testing the proportional hazards assumption

To check the Proportion Hazard assumption of Cox model, we use the scaled Schoenfeld residuals of the final fitted model. We plot these Schoenfeld residuals with all the covariates in the final model, namely vehicle type, premium value, assured sum, lease type, and age range. Since there are no apparent patterns in any of the plots in Figure 4, we can assume that the proportional hazard assumption holds.

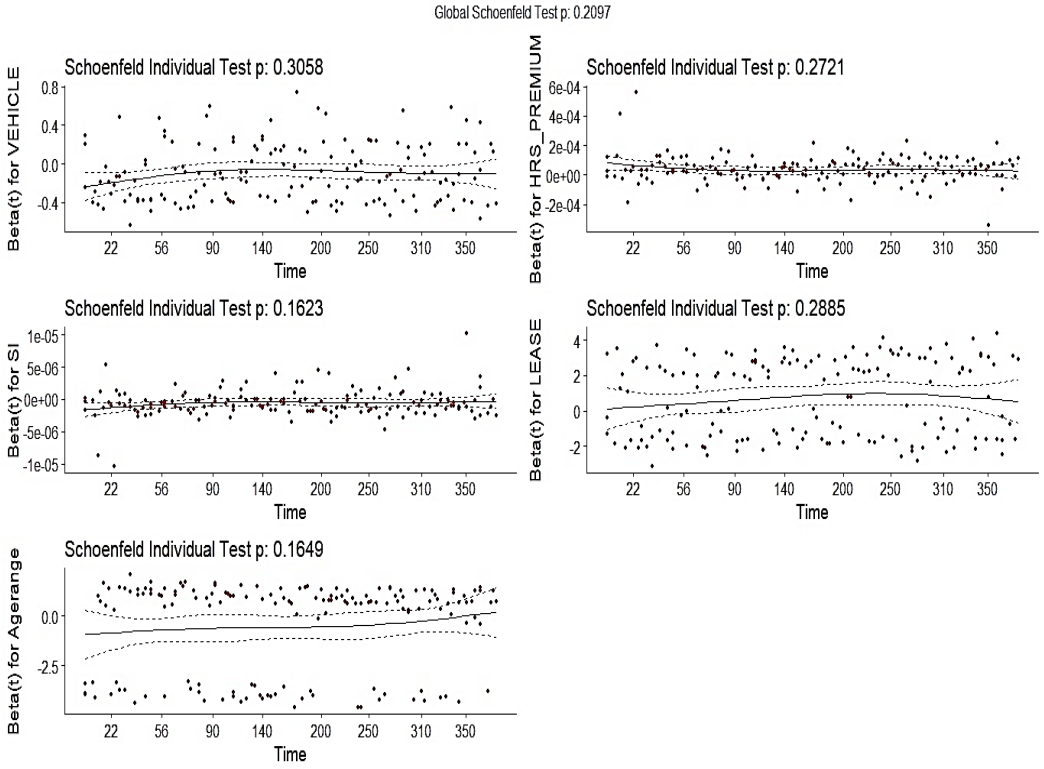

Fig 4: Schoenfeld residual graphs for vehicle type, premium value, assured sum, lease type, and age range

In Figure 4, the solid line represents a smooth spline fit to the plot and the dashed line represents a ± 2 standard-error band around the fit. Since there is no pattern with time, the assumption of proportional hazards appears to be satisfied by the above mentioned covariates.

Checking for influential observations

We use the DFBETA values to identify influential observations or outliers. Here DFBETA measures the difference in each parameter estimate with and without the influential point (Montgomery *et al*. 2006). Figure 5 displays these DFBETA values from the final fitted model. We use the usual cutoff value,$|DFBETA| > 2/\sqrt{n}$ to identify an influential observation. Here n is the sample data value in the training dataset which is used for training the Cox model. This cutoff is 0.08 for our study. There are only a few DFBETA values that correspond to large premium values and assured sum. But we can neglect these points in the dataset.

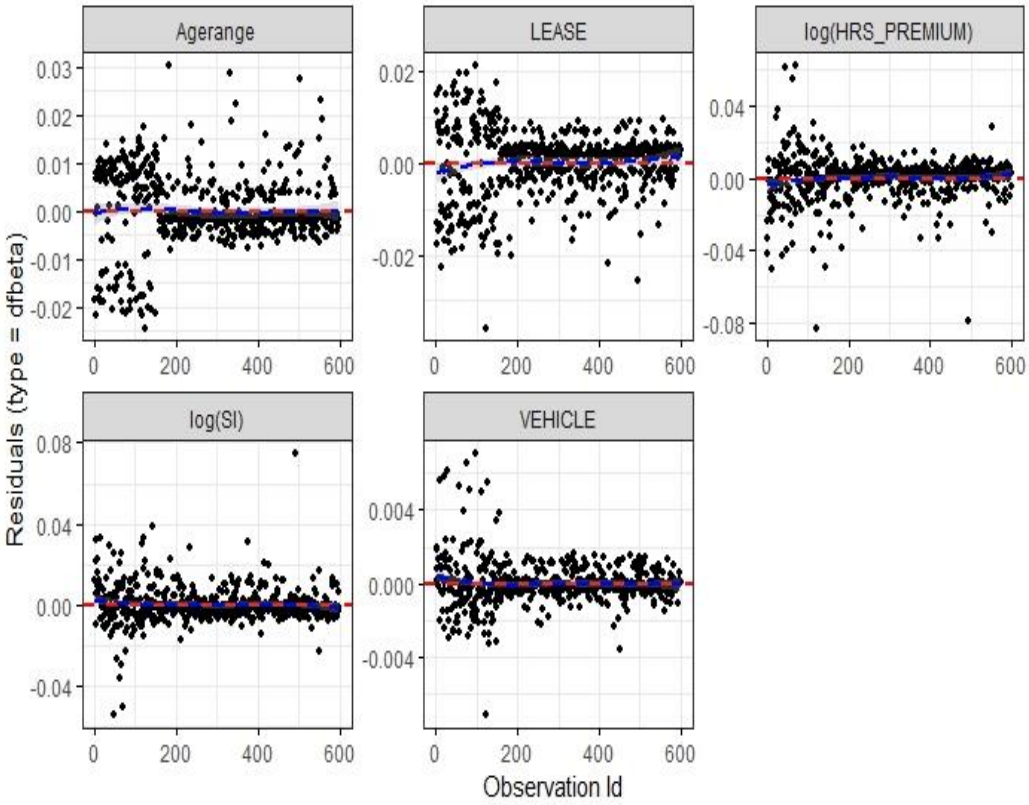

Fig 5: **Index plots of DFBETA values for the fitted Cox regression versus age range, premium, lease type, assured sum, and vehicle type.**

Detecting nonlinearity

Nonlinearity, an incorrectly specified functional form in the parametric part of the model, is a potential problem in Cox regression like in linear and generalized linear models. The martingale residuals may be plotted against covariates to detect nonlinearity. For nonlinear assumption for Cox's model, continuous type covariates should be linear. In our study, this assumption is checked only for the predictors, namely, assured sum and premium value. Figure 6 displays these two residual plots which are clearly not linear.

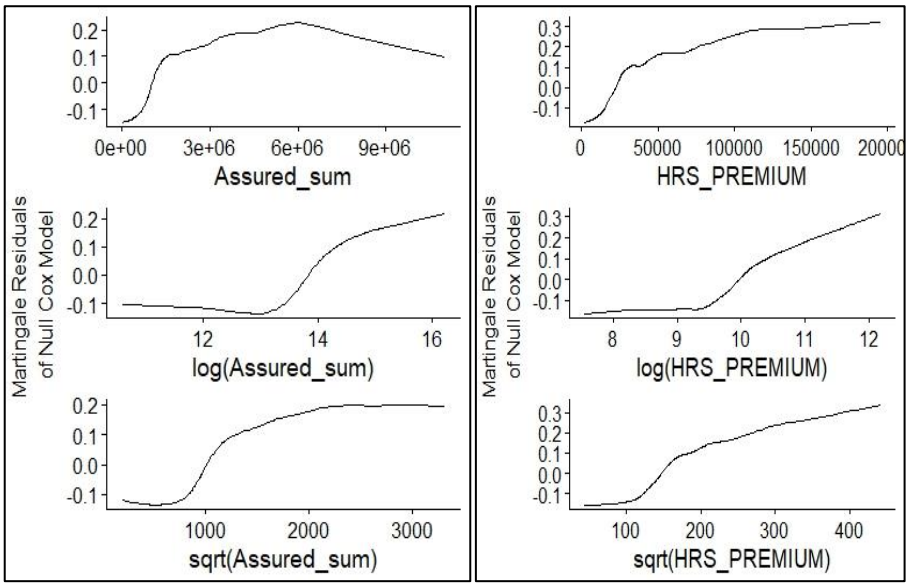

Fig 6: **Martingale residual graph for covariates, Assured sum and Premium.**

In Figure 6, we plot the martingale residuals against the two covariates, Assured-sum and Premium. We also fit the LOWESS smooth based on a span 0.2. Here, null Cox model means the model with all continuous type variables. To correct this nonlinear violation, we use logarithm transformations of both above variables in the fit. The results are displayed in Table 4. All predictors in this model fit are significant at the 5% significant threshold and all residual plots seem to be good. But we do not report these plots here because they are very similar to the above residual plots. Therefore, we can treat this fitted model as our final model for the motor vehicle insurance data. So we can use this model for future predictions.

Table 4: Final fitted Cox model

Variable name	β (estimated coefficients)	SE (coef)	Wald	p-value
Vehicle type	- 0.07102	0.02604	- 2.728	0.006378
Log (Premium value)	1.28256	0.30479	4.208	2.58e-05
Log (Assured sum)	- 0.69764	0.24308	- 2.870	0.004105
Lease type	0.62199	0.17238	3.608	0.000308
Age range	- 0.47701	0.18084	- 2.638	0.008344

3.2 Classification of the risky customers using logistic regression model

In this subsection, we fit the logistic regression model to the response variable with the above nine predictors in the motor vehicle claim dataset. Here, we choose two levels,

namely, a risky customer and a not risky customer as our response variable and it is a qualitative variable with two levels. In this study, we treat a customer as a risk if his or her first claim size is above 30,000 during the contract duration; otherwise we consider him or her as not risky. To fit this model, we randomly select 500 data points from our original sample for training data and we treat the remaining 99 data points as the test data. Table 5 shows the coefficient estimates for a logistic regression model that uses all of the above predictor variables, but we present only significant variables at the 5% significant level. They are vehicle type, premium value, lease type, and age range.

Table 5: Parameter estimates in fitted Logistic regression model.

	Estimate	Std.	Error	Z value
(Intercept)	1.94E+00	5.55E-01	-3.498	0.000468
Premium value	1.41E-05	5.91E-06	2.386	0.017011
Lease	2.46E-01	3.91E-01	2.63	0.005287
Vehicle	-1.44E-01	5.31E-02	-2.718	0.006576
Age range	-8.04E-01	3.89E-01	-2.064	0.039042

Table 6: Confusion matrix for the test dataset

		True default status		
		No	Yes	Total
Predicted risk value	No	82	3	85
	Yes	13	1	14
	Total	95	4	99

Next, we use the fitted logistic regression model with five significant predictors, namely, vehicle type, premium value, lease type, and age range to predict the customers in the test dataset as a risky or not risky customer. The results are displayed in the following confusion matrix for the test dataset in Table 6.

For the test dataset, the logistic regression model makes 83 correct classifications out of 99 sample data and the percentage accuracy is 0.8384. This achieves high accuracy. But logistic regression model makes only 16 misclassifications, and the percentage misclassifications is 0.1616.

The final goal of any insurance is to create profitable portfolio by assessing risk factors. As we mentioned, many of the research work conducted earlier related to motor insurance contracts analyzed the claim sizes with Cox proportional hazard models without conducting a proper residual analysis. In our study, we analyzed the time of the first claim of motor vehicle insurance contracts by using Cox's proportional hazard model. We identified five predictors as the most influential to the time of the first claim. These findings will help the company to take the decisions regarding the premium size of the contract for the different age groups, for example, according to their risk. The survival estimates of the time of the first claim obtained by using

Kaplan-Meier estimator will give the idea of the distribution of these claim time data. Since one contract may have more than one claim in its contract period, in future studies, we can analyze this type of data by treating them as recurrent event data. By conducting a thorough residual analysis, we found the optimal transformation of some selected covariates. Finally, we obtained a simple way to identify the risky customers by applying the logistic regression model. In future studies, one can apply more sophisticated methods like linear or quadratic discriminant analysis and Bayesian classifier methods to high accuracy.

Supplementary Material

Supplementary material for this paper is available separately.

Acknowledgments

Two anonymous reviewers are acknowledged for valuable comments on the initial draft of the manuscript.

References

Oulidi A, Marion J-M, Ganachaud H. 2010. Survival analysis methods in insurance applications in car insurance contracts. (accessed at https://studylib.net/doc/8687251/survival-analysis-methods-in-insurance-applications-in-car) (published online).

Cox DR. 1972. Regression models and life-tables. *Journal of Royal Statistical Society* B 34: 187-200.

De Mel Withanage Ajith Raveendra. 2014. On some inferential problems with recurrent event models. Doctoral Dissertation, Missouri University of Science and Technology (accessed at http://scholarmine.mst.edu/doctoraldissertations/2340.

Anna E. 2017. Variable selection for the Cox proportional hazards model. Master thesis. UMEA University (accessed at http://www.divaportal.org/smash/get/diva2:1067479/FULLTEXT01.pdf).

Fleming, Thomas R, and Harrington, David F. 1991. *Counting processes and Survival Analysis*. John Wiley & Sons Inc, New York.

Kaplan EL, Meier P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 153 (282): 457-481.

Montgomery DC, Peck EA, Vining GG. 2006. *Introduction to linear regression analysis*, fourth edition. John Wiley & Sons Inc, New York.

Nelson W. 1972. Theory and application of hazard plotting for censored failure data. *Technometrics* 14: 945-966.

Terry MT, Patricia MG. 1990. *Statistics for Biology and Health.* New York: Springer Verlag.