International Journal of Statistics and Probability; Vol. 6, No. 6; November 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Analyzing the Customer Attrition using Survival Techniques

Hasanthika, N. H. E.[1], & Jayasekara, L. A. L. W.[2]

[1]The Department of Mathematics, The Open University of Sri Lanka, Sri Lanka

[2]The Department of Mathematics, University of Ruhuna, Sri Lanka

Correspondence: Hasanthika N. H. E., The Department of Mathematics, Faculty of natural Sciences, The Open University of Sri Lanka, Sri Lanka, E-mail: erandihasanthika@yahoo.com

Received: August 9, 2017 Accepted: September 13, 2017 Online Published: September 27, 2017

doi:10.5539/ijsp.v6n6p85 URL: https://doi.org/10.5539/ijsp.v6n6p85

Abstract

Survival analysis techniques are used to study the amount of time between entry into observation and a subsequent event in estimating insurance attrition. Retention has always been a worldwide concern. A study is carried out on the profile of the policyholder and policies that produces better persistency based on one of the Sri Lanka experience using the nonparametric analysis (e.g. Kaplan-Meier estimator and life table analysis) and Cox regression model available through SPSS Statistics 20. This paper uses the survival model to evaluate the impact of covariates on the survival curves over a period of time. Newly opened life policies were considered during the period of 1st of January 2013 and 30th of June 2014 and our study period was end at 30th of June 2016. Survival analysis techniques can take into account for dealing with time-dependent variables and can help researchers to understand how insurance attrition impacts to the economic environment. Instead, the survival model provides much more information to the management and the people who deal with policies than what the regression model can offer.

Keywords: attrition, survival analysis, lapsation, Kaplan-Meier curves, proportional hazard model

1. Introduction

Survival analysis pertains to a statistical approach designed to allow the study of time between entry into observation and a subsequent event and is also named as time to event analysis. Kaplan and Meier (1958) are pioneers of the study of survival analysis and proposed to estimate survival functions from lifetime analysis and Kaplan-Meier curves using a series of horizontal steps of declining magnitude. Cox introduced the proportional hazard model in 1972, which describes the statistical relationships between a set of covariates and the survival function. Cox's proportional hazard model significantly improves the applicability of survival analysis, hence it is considered as a milestone in survival techniques. Survival analysis plays an important role in biomedical sciences where the event of interest was death and the dependent variable of study is often time to death. However, these techniques are also widely used in the engineering, social and economic sciences, and events or outcomes are defined by time to event data (Hosmer and Lemeshow (1999)). Examples include time until onset of disease, time until stock market crash, time until equipment failure, Time until tumor recurrence, time until cardiovascular death after some treatment intervention, time until designation of an employee and so on. Survival analysis techniques allow for us to start without all experimental units enrolled into study and to end before all experimental units have experienced an event. This is extremely important because even in the most well developed studies, there will be subjects who decide to quit participating, who move too far away from follow up, who have a lapsation (Bull and Spiegelhalter (1997)) from some unrelated event, or will simply not have an event before the end of the study period. Instead, considering the censored observations (Efron (1977) and Cnaan et al (1989)) is must because it enables researchers to analyze incomplete data due to delayed entry or withdrawal from the study. That means censored observations are observations which have the incomplete survival time (Lagakos (1979)). This is most important in permitting each experimental unit to contribute all of the possible information to the model for the time measured which allows the researcher to observe the unit. In this paper we focus on insurance retention and attrition (Wu and Lin (2009)). The event of interest is the policy attrition (either through end-term non-renewal or mid-term cancellation).

2. Survival Analysis and the Proportional Hazard Model

In survival analysis, time is always a continuous random variable and then the probability of an event at a single point of a continuous distribution is zero. It is important to define the probability of the events over a distribution by graphing the distribution of event times. The reader should start with the same fundamental tools of survival analysis, such as the more detailed description of the probability density function (pdf), the cumulative distribution function (cdf), the hazard

function, and the survival function, which can be found in any intermediate level statistical textbooks and it is important to note that one-one relationship that these four functions possess. Instead, the most important concepts in survival analysis are survival and hazard functions. To ensure that, further readings will be assist you with better understand in survival analysis techniques.

Suppose that the continuous probability distribution of a random variable, such as time, in survival analysis for a particular subject in study period is described by a cumulative distribution function. The cumulative distribution function (cdf) of a random variable, denoted *F (t)*, is defined by

$$F(t) = \text{Pr}ob(T < t)$$

The probability density function (pdf) or event density is also very useful in describing the continuous probability distribution of a random variable. The pdf of a random variable *T*, denoted *f (t)*, is defined by

$$dF(t)\, f(t) = dt$$

and every continuous random variable has its own density function. The density function represents the rate of attrition per unit of time. Let *T* denote the time until attrition occurs.

Then the survival function, *S(t)* takes on the following form:

$$S(t) = \text{Pr}ob(T \geq t) \text{ , where } t \geq 0$$

$$= 1 - F(t)$$

The survival function is the probability that the attrition occurs later than some specified value *t* that we choose. Because *S(t)* has the probability, ranges from 0 to 1 and then it is defined as *S(0) = 1* and as *t* approaches ∞, *S(t)* approaches 0. The hazard function *h(t)* is the ratio of the density function to the survival ($f(t)/S(t)$). In actuarial science, the $S(t)$ is the hazard function is often called the force of mortality. In this paper, we are considering policy attrition as an event of interest. Therefore, the hazard function gives a measure of the tendency of attrition: the greater the value of the hazard function, the greater the probability of attrition. Instead, because hazard function attempts to quantify the instantaneous risk that an event will take place at time *t* given that a particular subject survived to time *t*, it seems to be more intuitive to use the hazard function than the pdf in survival analysis. Exponential and Weibull distributions are the most popular survival distributions. The survival and density functions associated with the exponential distribution are $S(t) = e^{-\lambda t}$ and $f(t) = \lambda e^{-\lambda t}$, respectively. The hazard function for the exponential distribution is constant, $h(t) = \lambda$. The survival and density functions associated with the Weibull distribution are $e^{-\beta t^{\alpha}}$ and $f(t) = \alpha\beta^{\alpha-1}e^{-\beta t^{\alpha}}$ respectively. The hazard function of the Weibull distribution is $h(t) = \alpha\beta^{\alpha-1}$. When $\alpha > 1$, hazard rate is increasing over time. When $\alpha < 1$, hazard rate is decreasing over time.

When $\alpha = 0$, the hazard rate is constant over time.

In survival analysis, time to event for real applications is often not known because the event of interest may not occur before the end of study. In other words, the study is unable to wait for an event from a subject before the considered study period ends, and then this is called "right censoring" (Klein et al (1997)). In other words, a right censored subject's time terminates before the outcome of interest is observed. For example, in the context of analyzing the insurance policy attrition, if a policy is still active with an insurance company when the study ends, the data is right-censored. According to right censoring, time duration is only partially known above a given value. Survival analysis provides powerful tools through SPSS 20 to utilize this partial information without introducing statistical bias.

Kaplan-Meier Method

Kaplan-Meier method or one-sample nonparametric method (Kaplan et al (1958)) is used to estimate survival functions from lifetime data using a series of horizontal steps of declining magnitude. Because Kaplan-Meier method does not require any mathematical assumptions in obtaining hazard function or proportional hazard this is most interesting technique in comparing survival curves. In the context of Kaplan-Meier method, covariates within variables are not most important and it mainly deals with categorical predictors or grouped continuous variables. For the better understanding of failure events, the survival function is plotted as a stepwise reduction plot. In here, continuous variables are unable to consider directly, hence time-dependent variables will not be allowed. Because Kaplan Meier method treats censored data well, particularly right-censoring data, even today it is very important in survival analysis although these limitations exist.

Cox Proportional Hazards Model

In survival analysis, one model that is able to apply to determine which combination of potential explanatory variables affects the shape of the hazard function and to obtain an estimate of the hazard function for a particular study is the

proportional hazard model proposed by Cox (1972), which is also known as Cox regression model. There are no assumptions with Cox model in the nature or shape of the hazard function. The model can also take advantage of explanatory covariates on survival times. Cox regression model (i.e. the proportional hazards model) is most widely used as method for the analysis of censored data (Van Den Poel et al (2004)). The model equation is usually written as

$$h(t|x_t) = h_0(t)e^{\beta' x_t} \tag{2.1}$$

Where $h_o(t)$ is called the baseline hazard; it is called the hazard function when there are no covariate impacts. Where $x_t = (x_{1t}, x_{2t}, x_{3t}, ..., x_{kt})$, $\beta = (\beta_1, \beta_2, ..., \beta_k)$, where k is the total number of covariates, β_j is the constant proportional effect of x_j. Dividing both sides of Equation (2.1) by $h_0(t)$ and then taking the natural logarithm of both sides, we can obtain a linear transformation of the model:

$$log\left\{\frac{h(t \mid xt)}{h0\,(t)}\right\} = \beta' x_t \tag{2.2}$$

To estimate the β, Cox (1972, 1975) introduced the partial likelihood function. The statistical estimation of β has been studied expansively by Helsen and Schmittlein and has been obtained the semi-parametric partial maximum likelihood method as one of the popular numerical solutions. Suppose that a particular policy holder (say A) leaves the insurer at duration time t and then a number of other policies are at risk at t. Of all those policies at risk, policy A is the one that actually experienced the event (i.e. attrition) at t. Then the partial likelihood happened to the policy holder A within this duration time of t is given by the following equation:

$$L(\beta) = \frac{h_i(t)}{\sum_{j\epsilon R_i} h_j(t)} = \frac{e^{\beta' x_{i,t}}}{\sum_{j\epsilon R_i} e^{\beta' x_{j,t}}}$$

where R_t represents the risk set at t . The partial likelihood estimate of β can be obtained by maximizing this product; $\prod_{i=1}^{n}\left[e^{\beta' x_{i,t}} \Big/ \sum_{j\epsilon R_i} e^{\beta' x_{j,t}}\right]$ over observed n distinct ordered survival times.

3. Data Analysis

A particular insurance company wishes to evaluate survival time of its policies busing a follow-up study. Subjects were enrolled in the study from January 1, 2013 to June 30, 2014. The study period ended on June 30, 2016. Since the subject entered the study at different times over a one and half year period, the maximum possible follow-up time is different for each study participant. Policies A and C entered into the study at March 5, 2013 and June 10, 2013. Policies B and D entered into the study at February 01, 2014 and January 01, 2013. Policy A did not renew for the next term and policy C cancelled the term as shown in Figure 1. Policies B and D were inforce at the end of the study. There were a total of 158 policies at the beginning of the study. When we graph the estimator for survivorship function it clearly status that how to represent the actual values. For example, consider the first interval [0,3), where the value of the estimated survivorship function is reported in Table 1 as 0.68. Other intervals would be represented in a similar manner.

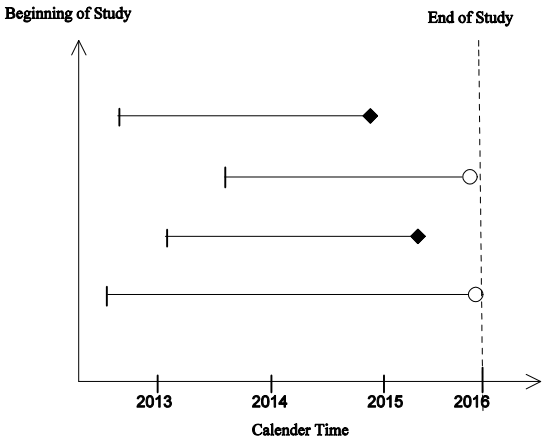

Figure 1. Line plot in calendar time for four subjects in the study

Table 1. Life table analysis for the view of retention and attrition

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number of Terminal Events	Cumulative Proportion Surviving at End of Interval
[0,3)	158	0	51	.68
[3,6)	107	0	16	.58
[6,9)	91	0	7	.53
[9,12)	84	0	6	.49
[12,15)	78	0	6	.46
[15,18)	72	0	8	.41
[18,21)	64	0	5	.37
[21,24)	59	0	2	.36
[24,27)	57	20	6	.31
[27,30)	31	7	2	.29
[30,33)	22	5	3	.25
[33,36)	14	3	0	.25
[36,39)	11	8	0	.25
[39,42)	3	3	0	.25

To detect the attrition and retention patterns by individual subjects, we check the Kaplan Meier estimate for all subjects in the study as in Figure 2. It noticeably represents the censored observations by "cross" sign with the time period. The cumulative survival probabilities are derived by assuming other variables are at their average values. Figure 3 represents the baseline survival curves for male and female policy holders in the run-time. In Figures 2, the green lines are randomly above the blue lines, which demonstrate that male policy holders are more likely to renew their insurance policies than the female policy holders. But in advance, Wilcoxon test is used to compare survival distribution among groups, with the test statistic based on differences in group mean scores. Table 2 shows the significance value of the test is greater than 0.05 and then we conclude that the survival curves are similar across the gender groups.

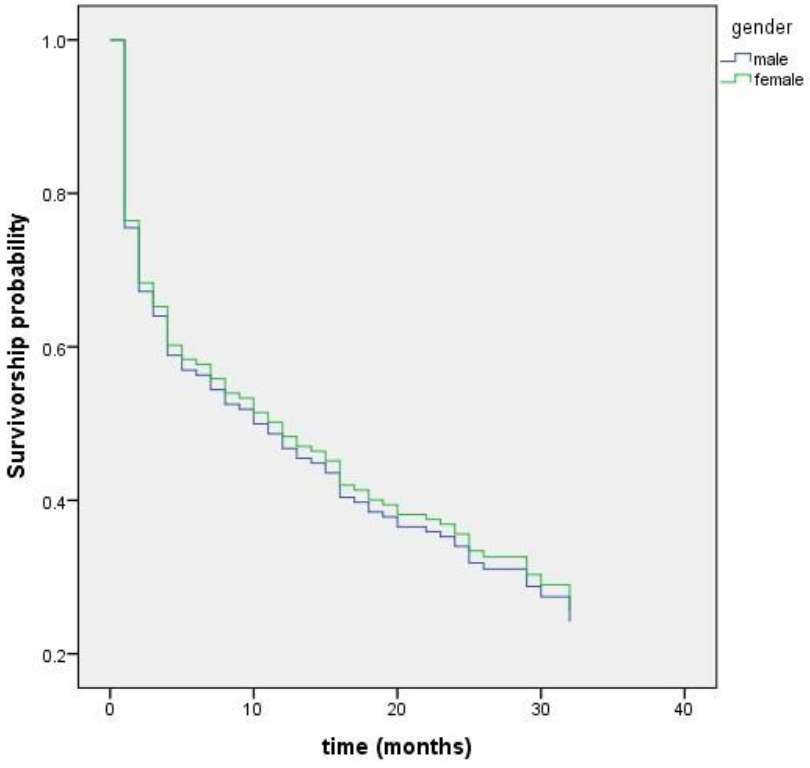

Figure 2. Estimated survivorship function defined by gender

Table 2. Use Wilcoxon test to compare survival distributions of gender

Wilcoxon (Gehan) Statistic	Degrees of freedom	p-value
.002	1	.967

Table 3 provides overall comparison tests of the equality of survival times across gender groups. Since the significance values of the tests are all greater than 0.05, there is no statistically significant difference between two gender groups in survival time. As in Table 4, there is a lot of overlap in the confidence intervals; it is unlikely that there is much difference in the "average" survival times. If confidence intervals do not overlap between levels, differences in effect on time to event can be inferred.

Table 3. Test statistics, degrees of freedom and p-values for the equality of the survivorship functions among the two gender groups in the study.

Statistic	Chi-Square	Degrees of freedom	p-value
Log Rank (Mantel-Cox)	.004	1	.952
Breslow (Generalized Wilcoxon)	.002	1	.967
Tarone-Ware	.006	1	.938

Table 4. Estimator, standard error and 95% confidence interval of Means and Medians for Survival Time

Gender		Mean			Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
male	17.060	1.907	13.323	20.797	12.000	3.765	4.621	19.379
female	16.675	1.729	13.286	20.065	10.000	3.751	2.647	17.353
overall	16.817	1.280	14.308	19.326	11.000	2.957	5.204	16.796

The nature of the Cox proportional hazard model allows us to approximate the coefficients relate to the covariates in a study. Table 8 displays the coefficients of four selected variables from survival analysis: age, gender, policy type and mode of payment. It reports the coefficients of age and gender is consistent with insurance attrition since the significance values are less than p-value (0.05). But, the value of the coefficient of age is very small. Other two variables; mode of payment and policy type are not consistent as the significance values are higher than 0.05. The sign of the coefficient for age is negative, implying that older is more likely to stay with an insurer. The sign of the coefficient for gender is positive, implying that it will drive up the probability of attrition.

Table 5. Parameter estimates using survival analysis

Variable name	B (estimated coefficient)	SE	Wald	df	p-value
Gender	.533	.203	6.860	1	.009
Age	-.010	.001	101.162	1	.000
Type of policy	.019	.027	.470	1	.493
Mode of payment	.269	.391	.471	1	.493

Table 6. Adjusted parameter estimates using survival analysis

Variable name	B (estimated coefficient)	SE	Wald	df	p-value
Gender	.564	.202	7.848	1	.005
Age	-010	.001	103.759	1	.000

4. Conclusion

Actually, survival analysis provides a new perspective on the persistency problem beyond the information about the duration of holding a policy. Insurance attrition minimizes the size of sales or the top line of an insurance company. Therefore, it should be carefully scrutinized to develop the methods for retention policies. Instead the profit gaining from policy holders, retention tendency of the customer should be understand deeply to estimate the customer accurately. In this paper, survival analysis has been applied as an alternative approach to the other regressions to analyze insurance attrition using the tools available through SPSS Statistics 20. Survival analysis uses the continuous time as the response variable and is responsible for answering not only whether but also when a policy will leave. In the expiration month, a significant number of policies have no renewal process and this scenario will happen in further months. Survival analysis is able to model the cancellation and nonrenewal sequentially and capture this seasonality of attrition well. By analyzing the life table and the survivorship probability plot, more information can be gathered to policyholder, product and agent that provide better persistency for the insurance industry. Cox proportional hazard model focuses upon the variables with higher impact of creating a validate model because of time-dependent macroeconomic variables affect to insurance retention and attrition. Thus, agents are able to identify which variables related to the better persistency of policies within their units and management as well.

References

Allison, P. D. (1995). *Survival Analysis Using the SAS System*. SAS Institute.

Anderson, K. M. (1991). A non-proportional hazards Weibull accelerated failure time regression model. *Biometrics, 47*, 281–288. https://doi.org/10.2307/2532512

Bull, K., & Spiegelhalter, D. J. (1997). Tutorial in Biostatistics Survival Analysis in Observational Studies. *Statistics in Medicine, 16*, 1041–1074. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<1041::AID-SIM506>3.0.CO;2-F

Cnaan, A., & Ryan, L. (1989). Survival Analysis in Natural History Studies of Disease. *Statistics in Medicine, 8*, 1255–1268. https://doi.org/10.1002/sim.4780081009

Cox, D. R. (1972). Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society Series B, 34*, 187–220.

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association, 72*, 557-565. https://doi.org/10.1080/01621459.1977.10480613

Fleming, T. R. & Lin, D. Y. (2000). Survival Analysis in Clinical Trials: Past Developments and Future Directions. *Biometrics, 56*, 971–983. https://doi.org/10.1111/j.0006-341X.2000.0971.x

Hosmer, D., & Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data.* Wiley-Inter science.

Kaplan, E. L., & Meier, R. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association, 53*, 457–81. https://doi.org/10.1080/01621459.1958.10501452

Klein, J. P., & Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*. Springer. https://doi.org/10.1007/978-1-4757-2728-9

Lagakos, S. W. (1979). General Right Censoring and Its Impact on the Analysis of Survival Data. *Biometrics*, *35*, 139–156. https://doi.org/10.2307/2529941

Scheike, T. H., & Martinussen, T. (2006). *Dynamic Regression Models for Survival Data.* Springer, New York.

Tanser, J. (2010). *Conversion and Retention Modeling*. Ratemaking and Product Management Seminar, Casualty Actuarial Society, Chicago.

Van den Poel, D., & Lariviere, B. (2004). Customer Attrition Analysis for Financial Services using Proportional Hazard Models. *European Journal of Operational Research, 157*(1), 196–217. https://doi.org/10.1016/S0377-2217(03)00069-9

Wu, C. S. P., & Lin, H. (2009). *Large Scale Analysis of Persistency and Renewal Discounts for Property and Casualty Insurance*. CAS E-Forum, Winter, 396–408.