EMPIRICAL STUDY OF TIME TO LAPSE FROM LIFE INSURANCE PARTICIPANTS USING *SURVIVAL ANALYSIS*

Rizky Pradian Sukma, S.Si, Dr. Yogo Purwono FRM

1. Faculty of Economics and Business, Depok, West Java, 16424, Indonesia
2. Faculty of Economics and Business, Depok, West Java, 16424, Indonesia

E-mail: rizky_pradian@yahoo.co.id

ABSTRACT

Every life insurance company tries to maintain the degree of persistence of the policy. If many customers drop out of contracts in the middle of the insurance period (lapse), it will influence the number of premiums and insurance reserves that will be formulated. This research study considers the estimated survival opportunities and lapses rates for term life insurance products in the agency business line at PT Asuransi XYZ with an observation period of January 1st, 2013 to December 2017 using truncated data and right sensors. The method used is the nonparametric method by Kaplan Meier and Nelson Aalen and the parametric method of Weibull Distribution by considering the variables of gender, age, type of premium payment, premiums, duration of premium payments, and branches. The selection of the best model uses the Mean Standard Error and AIC. The results of the analysis obtained by the two methods are that the lapse level will decrease until the end of the period.

Keywords: Kaplan Meier, Lapse, Nelson Aalen, Survival, Weibull.

Introduction

According to Republic of Indonesia Law Number 40 of 2014 about Insurance, life insurance business is a business that carries out risk management services which provide payments to policyholders, insured, or other parties who are entitled in the event the insured dies or remains alive, or other payments to policyholders, insured, or other parties who are entitled at a particular time regulated in the agreement, the amount of which has been determined and / or based on the results of the management of funds. The agreement between the life insurance company and the policyholders / insured is stated in the form of an insurance policy. The insurance policy must include the minimum provisions regarding the validity

period, description of the agreed benefits, the way to pay premiums or contributions, grace period for premium or contribution payments, the exchange rate used for insurance policies with foreign currencies if premium payments or contributions and benefits are linked to the rupiah currency, time admitted as the time of receipt of premium payments or contributions, company policy stipulated if the premium payment or contribution exceeds the agreed time period, the period when the company cannot review the validity of the insurance contract (incontestable period) on the product long term insurance, et cetera.

Lapse rates in an insurance product can affect the making of actuarial assumptions. One of the applications of actuarial science in the insurance discipline is in the calculation of premium pricing and calculation of future reserves of insurance policies. In deciding premium prices, an actuary must have assumptions such as assumptions on the rate of return on investment and costs, movements in the number of policies and claim rates, assumptions about liabilities/reserves, and assumptions of mix point models. While to make a reserve valuation, the assumptions used are the best estimation assumption of the provision for risk margins. Besides, the interest rate assumption also uses the assumptions of mortality/morbidity/other claims, cost assumptions and lapse assumptions.

Based on Final Draft of Actuary Practice Standards - Indonesian Actuarial Association in Technical Guideline 3, the backup method is based on gross premium valuation. The assumption of lapses in the calculation of reserve valuation is determined by considering factors such as the type of product, age of the insured, duration of policy and method of premium payment and premium payment period. Besides that, are the risk selection method, premium payment status, amount of sum insured, and premium amount. Then regarding marketing, distribution channels and interest rate factors, cash values, as well as bonuses, taxes to surrender fees and others.

In this research study, the calculation of chance estimates of lapse is based on survival analysis using parametric and nonparametric methods. Nonparametric methods only discuss individuals in groups and survival time, so a statistical approach that can explain the relationship of variables that affect individuals in survival time is needed. The parametric method is a survival model with survival time that follows specific distribution assumptions. The advantage of the parametric model is that survival time follows a certain distribution and can also forecast the time of an event up to the period of an event occurring in the observation data. Whereas for the nonparametric method, the Kaplan Meier and Nelson Aalen Functions

models are used to estimating the levels of lapse appropriately based on historical data between the life insurance company business with the help of nonparametric methods.

This study will analyze the level of lapse and survival in PT Asuransi XYZ company that occur in term insurance products in the agency business line. It is a significant growth in 2017. Based on data reported by the Financial Services Authority (OJK) through the 2017 Insurance Statistics, the number of insured people who went to PT Asuransi XYZ increased by 51.13% while the life insurance industry decreased by 90.87%. However, Sum Insured (UP) which experiences lapse at PT Asuransi XYZ increased by 202.95% while the life insurance industry decreased 47.65%. Looking at these problems, the researchers consider this to be the focus for research using survival analysis by considering several variables such as gender, age, type of payment, amount of premium, duration of payment and branches.

Survival method

Survival analysis is a method that is associated with time, starting from the time when it starts up to the occurrence of a special event. Survival functions can be used to perform calculations on the chances of survival at a specific time. In the survival function, there are several notations such as *x* which state that as age, where x ≥0, the probability of death occurs at age is greater than x. For example. *S (t)* is a survival function defined as follows:

$$S(t) = P(T \geq t) \tag{1}$$

Based on the definition of the cumulative distribution function *F (t)* of *T*, the survival function can be expressed by,

$$S(t) = 1 - P(T \leq t)$$

Hazard function *h (t)* is expressed as the speed of an individual will experience an event in the time interval from *t* to $t + \Delta t$ with the condition that the individual still lives up to time *t*, that can be expressed by the following equation:

$$h(t) = \frac{f(t)}{S(t)} \tag{2}$$

Kaplan Meier

Kaplan Meier is a method used to estimate *S (t)* or often also called Product-Limit Estimator. The Kaplan Meier method is one of the descriptive methods of survival analysis as well as Life Table analysis. Although Life Table method is a method developed for the first time in survival analysis, the Kaplan Meier method has some advantages in various conditions compared to the Life Table method.

Kaplan Meier analysis can produce a survival curve for a population and many statistics such as the median of survival time. The use of Kaplan Meier's analysis in statistical science is to analyze how a population will change according to the time/period that is running. Three reasons why an individual or product can evolve are stated below:

- The individual or product will die;
- Some of the population will leave the survey because they are repaired or cured;
- Alternatively, because the data is lost from observation (individuals move, studies are stopped, et cetera.)

The first type of data is data failure or data from an event, while the second and third data types are censored data.

The following are estimates from the Kaplan Meier method:

$$\widehat{S}(t) = \begin{cases} 1 & \text{jika } t < t_1 \\ \prod_{t_i \le t} \left(1 - \frac{d_i}{Y_i}\right) & \text{jika } t_i \le t \end{cases} \quad (3)$$

d_i is the number of events

Y_i is the number of individuals who have risks (number at risk).

Nelson Aalen

Nelson Aalen's method or in other words is empirical cumulative hazard function is a method proposed by Nelson in 1969 and Aalen in a thesis in 1972.

Estimator for cumulative hazard functions:

d_i is the number of events

$$\widehat{H}(t) = \begin{cases} 1 & \text{jika } t < t_1 \\ \sum_{t_i \le t} \left(\frac{d_i}{Y_i}\right) & \text{jika } t_i \le t \end{cases} \quad (4)$$

Y_i is the number of individuals who have risks (number at risk).

Stepwise Regression

Stepwise regression is one of the procedures for selecting the best set of predictor variables. The analysis approach follows the multivariate regression stages as follows (Fahrmeir et al, 2013; Keith, 2015):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + e \quad (5)$$

Where

Y = dependent variable

β_0 = regression constant

$\beta_1, \beta_2, .., \beta_n$ = regression coefficient

$X_1, X_2, .., X_n$ = independent variable

e = estimated error (residual)

Stepwise regression can be described by fundamentel steps (algorithm) as follows (Hanke & Weihern, 2005):

1. Determination of the correlation matrix between the dependent variable Y to the independent variable.
2. The independent variable that has a correlation coefficient with the dependent variable is the first variable that goes into the regression equation
3. The next variable that enters equations is one variable (other than the one that was entered before) which has a significant contribution to the number of significant squares of the variables entered in the regression equation determined by the F test. The value of the statistic F that must be exceeded by the independent variable is called F to enter.
4. When additional variables are included in the equation, the individual contribution to the number of regression squares of the other variables that have been entered in the equation is calculated using the F test. If the F statistic is less than a value called F to remove, then the variable is omitted from the regression equation

5. Interpretation of the model obtained.

2.4.2 Weibull Distribution

One of the survival data in analyzing is the Weibull distribution. Weibull distribution has a density function with parameters λ>0 and p>0. The following is a function of the Weibull distribution:

$$f(x) = \lambda p x^{p-1} \exp(-\lambda x^{p}) ; x \geq 0 \quad (6)$$

The survival function of the Weibull distribution is

$$S(x) = \exp(-\lambda x^{p}) ; x \geq 0 \quad (7)$$

The hazard function of the Weibull distribution is

$$h(x) = \lambda p x^{p-1} \quad (8)$$

The parameter p is called a shape parameter because this parameter determines the change in the hazard curve as time x increases. if p>1, then the monotonous hazard function rises. If p<1, then the monotonous hazard curve decreases. Whereas for p=1, the hazard is constant and follows the special form of the exponential distribution with h (x) = λ. The shape parameter makes the Weibull distribution flexible for survival data.

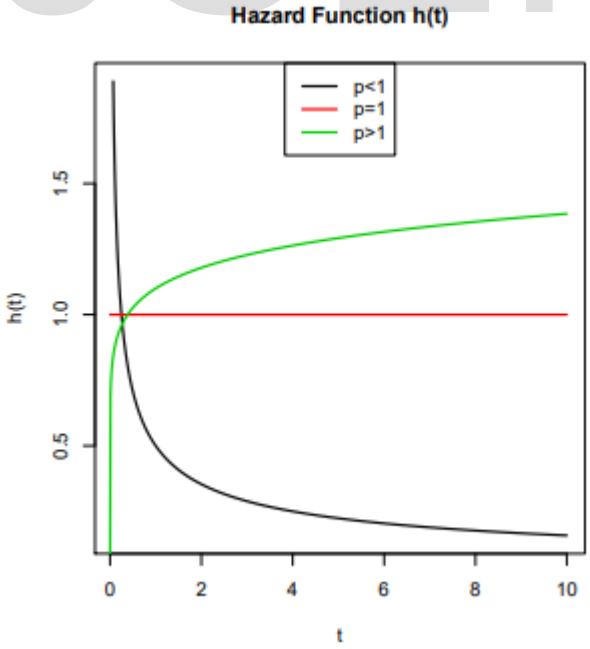


Figure 2.1 Weibull distribution hazard function

Method of Model Selection

The determination of the best model selection of non-parametric to estimate in this research used Standard Error (SE). Standard Error is used to measure various models of data samples in research object. Decent model usually has the smallest standard error. The calculation for standard error value is estimated by estimating sample data from population deviation data with the root from the number of sample data of research data.

$$SE = \frac{Std.dev}{\sqrt{n}} \tag{9}$$

$Std.dev$ is population deviation standard
n is the number of sample data

The comparison of the best model can be determined by using Akaike Information Criterion (AIC) calculation. The best model can be chosen if AIC value is the most minimum. The value of AIC as follow:

$$AIC = e^{\frac{2k}{n}} \frac{\sum_{i=1}^{n} \hat{u}_i{}^2}{n} \tag{10}$$

k is estimated parameter in used regression model
n is the number of observation
e is 2.718
u is the residual

Research Method

There were 2 (two) variables used in this research which were respond and explanatory variable. Respond variable was survival duration of customer during observation which was shown in month form, while explanatory variable used the gender, age of the customer and premium total paid by customer, premium payment type and duration of payment and branch.

Table 1. Observation data used for survival analysis

Time	Information	Observation
Time when customer has the opportunity to default premiums (for example: 1,10,60)	*0=sensor* *1=event*	Numeric code from gender, age, premium, premium payment type, duration of payment and branches.

Respond variable used in this research was the time of customer entry, the time of customer release and event which described about censored observed data (0 was censor) and data which had failure or and event was 1. For explanatory variable, the first was variable from Gender which was changed into 1 dummy variable of G_1 with 1 was men and 0 was women.

Table 2. Variable of dummy Gender

Variabel Penjelas	Keterangan	G_1
Gender	1. Men	1
	2. Women	0

The second Explanatory variable was age. It was assumed in 3 dummy variables. It was because the assumption of data grouping with age range of 31-40 years was assumed into a dummy variable became U_1, 41-50 years was assumed into dummy variable became U_2 and age above 50 years was assumed into variable of dummy U_3.

Table 3. Variable of age dummy

Explanatory Variable	Information	U_1	U_2	U_3
Age (Year)	1. 17 – 30	0	0	0
	2. 31 - 40	1	0	0
	3. 41 - 50	0	1	0
	4. Above 50	0	0	1

The third explanatory variable above was total premium paid by customer to the company. For the data of this premium income, it was divided into 4 groups assumed in 2 dummy variables. First group was premium payment of customer approximately IDR 2.3

millions to IDR 10 millions with variable of *dummy* P_1. The last was group 2 with range of premium payment more than IDR 10 millions with variable of *dummy* P_1.

Table 4. Variable of Premium dummy.

Explanatory Variable	**Information**	P_1	P_2
Premium (Rp)	1. IDR100.000,- IDR.2.500.000,-	0	0
	2. IDR.2.500.001, - IDR.10.000.000,-	1	0
	3. Above IDR.10.000.000,-	0	1

The fourth Explanatory variable was payment type conducted by customer which was assumed in 3 dummy variables. In this research, premium payment type was divided into monthly made as variable of dummy TP_1, payment in semester TP_2 and quarterly payment TP_3.

Table 5. Dummy variable of premium payment type

Explanatory Variable	**Information**	TP_1	TP_2	TP_3
Type of Premium payment	1. Monthly	1	0	0
	2. Quarterly	0	0	1
	3. Semester	0	1	0
	4. Annual	0	0	0

The fifth data variable was duration of premium payment. Coding conducted i this research was with payment duration from 61-120 in variable of *dummy* DP_1 and payment duration which was bigger than 120 was variable of *dummy* DP_2.

Table 6. Dummy variable of Payment duration

Explanatory Variable	**Information**	DP_1	DP_2

Premium Payment Duration (annual)	1. ≤ 60	0	0
	2. 60 - 120	1	0
	3. > 120	0	1

The sixth data variable was term life insurance product sales of branches. For branch with branch 2, it was assumed in variable of *dummy* C_1, branch 3 in variable of *dummy* C_2, branch 4 in variable of *dummy* C_3, branch 5 in variable of *dummy* C_4, branch 6 in variable of *dummy* C_5, branch 7 in variable of *dummy* C_6 , branch 8 in variable of *dummy* C_7.

Table 7. Variable of branch dummy

Explanatory Variable	**Information**	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Branch	1. Medan, Palembang, Jambi, Pekanbaru, Lampung dan Batam (Branch 1)	0	0	0	0	0	0	0
	2. Jakarta dan Bekasi (Branch 2)	1	0	0	0	0	0	0
	3. Cirebon, Bandung, Tasikmalaya dan Bogor (Branch 3)	0	1	0	0	0	0	0
	4. Purwokerto, Yogyakarta, Tegal dan Semarang (Branch 4)	0	0	1	0	0	0	0
	5. Surabaya, Jember, Denpasar, Gianyar, Sidoarjo dan Kediri (Branch 5)	0	0	0	1	0	0	0
	6. Makassar, Kendari, Manado, Palu, Banjarmasin, Balikpapan, Gorontalo dan Bau-Bau (Branch 6)	0	0	0	0	1	0	0
	7. Surakarta (Branch 7)	0	0	0	0	0	1	0

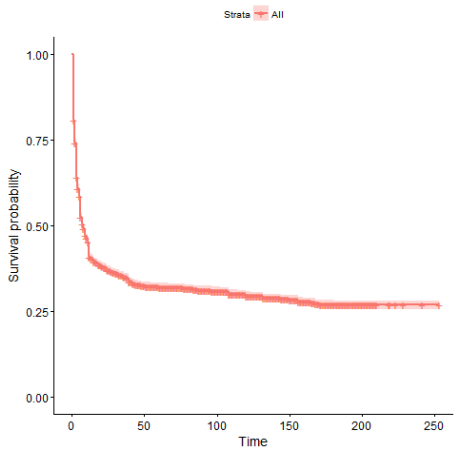
	8. Malang (Branch 8)	0	0	0	0	0	0	1

Result and Discussion

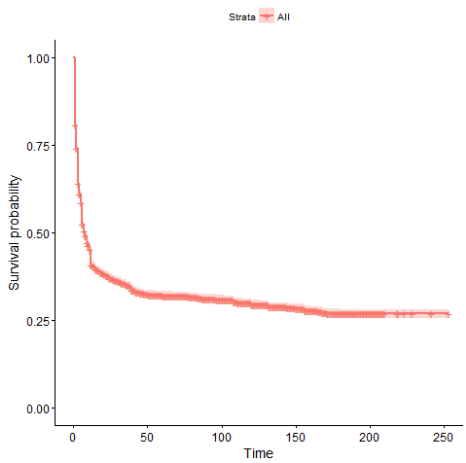
Table 8. Estimation of Customer Survival of Company Insurance XYZ in 2013-2017

Time	*n Risk*	*n Event*	*Censored*	**Kaplan Meier**		**Nelson Aalen**	
				survival	***std.err***	***survival***	***std.err***
1	7179	1378	410	0,808051	0,004648	0,808065	0,004648
…	…	…	…	…	…	…	…
12	3149	308	190	0,409849	0,006011	0,409893	0,006012
…	…	…	…	…	…	…	…
24	2646	39	144	0,372483	0,005911	0,37253	0,005912
…	…	…	…	…	…	…	…
36	1345	14	83	0,351084	0,005954	0,351134	0,005955
…	…	…	…	…	…	…	…
48	1719	9	147	0,325678	0,006156	0,325737	0,006157
…	…	…	…	…	…	…	…
60	2163	11	259	0,320997	0,006134	0,321056	0,006136
…	…	…	…	…	…	…	…
183	843	1	19	0,26941	0,006259	0,269483	0,006261

Kaplan Meier



Nelson Aalen



Based on table 8, estimation of survival opportunity by using Kaplan Meier method of $\hat{S}(183) = 0{,}26941$ with trust range of 95% has lower limit of 0,257418 and upper limit of

0,281961 and estimation of survival opportunity of Nelson Aalen method of $\hat{S}(183) = 0.269483$ with trust range of 95% had lower limit of 0.257487 and upper limit of 0.282037.

Table 9. Criteria of choosing Non-parametric Model

Metode	*Mean Standard Error*
Kaplan Meier	0.006055887
Nelson Aalen	0.006056889

Choosing the best model in non-parametric used Mean Standard Error. Form the table above, it can be concluded that Kaplan Meier was the best model. It is because MSE in Kaplan Meier is smaller than Nelson Aalen method.

Parametric Distribution Test

Tabel 10. Goodness-of-fit criteria

	Exponential	**Logistic**	**Normal**	**Loglogistic**	**Weibul**
Akaike's Information Criterion	153528.9934	167679.5674	167421.6006	154700.4237	151464.1782

Table 10 is a data distribution of lapse time. Based on AIC vale, it is obtained that the data is data with Wibull distribution that is taken from the smallest AIC value than other distributions. Furthermore, it will conduct survival model of Weibull parametric.

Stepwise Regression

The first step in stepwise selection was entering all explanatory variable, then it is obtained IC value, after that, eliminate explanatory variables one by one. The table below is the result of calculation using stepwise regression:

Table 11. The Entirety Stepwise Regression

Variable	DF	AIC
Sex	1	36994.7
<none>		36996.7
Premium	2	37027.3
Age	3	37031.8

Q_Payment	2	37178
Payment_Method_Gap	3	37693
Branch	7	38274,1

In this step, it is obtained that initial AIC value is 36996,7 and it can be obtained if it was reduced with AIC gender variable of 36994.7, AIC premium variable of 37027.3, AIC Age variable of 37031.8, AIC payment duration variable of 37178, AIC payment type variable of 38274.1. the next step is trying to add gender variable, the table below is the result:

Table 12. Stepwise Regression without Gender Variable

Variable	DF	AIC
<none>	1	36994.7
Sex		36996.7
Premium	2	37027.3
Age	3	37029.8
Q_Payment	2	37176.2
Payment_Method_Gap	3	37694.4
Branch	7	38279.2

Stepwise regression in wibull parametric survival model obtains the best model with the smallest AIC value of 36994.71. Variables entered in model are branch, payment method, total payment, age and premium and sex variable has been eliminated in weibul parametric model.

Weibull Parametric Model

After obtaining variable in model, the nest step was entering all variable in Weibull parametric model. This is the result obtained from R software:

Table 13. Weibull Parametric Model

Covariate		Mean	Coef	Exp (Coef)	Se(Coef)	Wald p
Branch	1	0.116	0	1	(reference)	0
	2	0.089	1.623	5.066	0.141	0
	3	0.271	1.912	6.768	0.125	0
	4	0.173	0.17	1.186	0.174	0.329
	5	0.125	1.251	3.495	0.14	0
	6	0.104	2.614	13.648	0.129	0
	7	0.066	0.959	2.609	0.18	0
	8	0.057	1.308	3.6999	0.159	0
Payment_method_gab	annually	0.421	0	1	(reference	
	Monthly	0.226	2.614	37.131	0.178	0
	Semester	0.083	1.176	3.241	0.104	0
	Quarterly	0.27	1.831	6.242	0.089	0
Q_Payment	<=60	0.701	0	1	(reference	
	.120	0.142	-1.46	0.232	0.153	0
	60-120	0.157	-1.581	0.206	0.149	0
Age	<=30	0.253	0	1	(reference)	
	>50	0.008	-2.082	0.125	0.501	0
	30-40	0.4	-0.063	0.939	0.036	0.081
	40-50	0.339	-0.076	0.927	0.037	0.038
Premium	<=2.5 millions	0.897	0	1	(reference)	
	>10 millions	0.009	-0.981	0.375	0.449	0.029
	2.5 - 10 millions	0.094	0.559	1.749	0.106	0

log(scale)	9,456	12783,734	0,247	0
log(shape)	-0,499	0,607	0,012	0

Events	4651
Total time at risk	290571
Max. log. likelihood	-18480
LR test statistic	3222,72
Degrees of freedom	17
Overall p-value	0

From the analysis, it was obtained the value of ratio test likelihood of 3222,72 with value $p<0.05$, it showed that at least, there was a coefficient which affected significantly toward the cause of lapse. Analysis result using weibull parametric survival with selection stepwise method gave information that variable with all different category toward the reference of alpha level 5% were variables of payment method, total payment and premium. Whereas branch and age variables had at least one category which did not give difference significantly toward the reference.

Conclusion

1. Based on the result obtained with Kaplan Meier and Nelson Aalem methods, there is survival opportunity estimation significantly affected by variables of gender, age, payment type, payment duration, premium and branch. Thus, estimation of survival opportunity from each variable obtains almost the same, from the begining of period until the end of period.
2. In PT Asuransi XYZ, estimation of lapse level of term life insurance product, it obtains from 2 ways of calculation which are using parametric and non-parametric methods. It generates directly proportional lapse level estimation that both ways decrease.
3. Based on analysis and the result, it can be concluded that Weibull distribution is optimum parametric model through the smallest AIC value, whereas Kaplan Meier is non-parametric model through value criteria of the smallest standard error rather than Nelson Aalen.

REFERENCE

Beenstock, M., G. Dickinson dan S. Khajuria. 1986. The Determinants of Life Premiums: An International Cross-Section Analysis 1970-1981. Insurance Mathematics and Economics. Vol. 5, No. 4, 261-270.

Barsotti Flavia, Milhaud Xavier, Salhi Yahia. (2016). *Lapse* risk in insurance: correlation and Contagion effects among policyholders' behaviors. Insurance :Mathematics and Economics.France

Emiliano A. Valdez, Jeyaraj Vadiveloo, Ushani Dias (2014) Life insurance policy termination and survivorship USA Mathematics and Economics 58 (2014) 138–149

Fitriani, Sri Wahyuningsih, dan Yuki Novia Nasution. (2016) Penggunaan Metode Nonparametrik Untuk Membandingkan Fungsi *Survival* Pada Uji Gehan, Cox Mantel, Logrank, Dan Cox F. Kalimantan Timur Volume 7 Nomor 2

Kitty J. Jager, Paul C. van Dijk, Carmine Zoccali and Friedo W. Dekker. (2008) The analysis of *survival* data: the Kaplan–Meier method. Amsterdam

Klein P. Jhon, Moeschberger L. Melvin.(2003). *Survival* analysis Techniques for Censored and Truncated Data (second edition) .USA

Klugman, Stuart, Panjer, Willmot. (2012). Loss Models From Data to Decisions : fourth edition. USA: Wiley Series

Moore F. Dirk. (2016). Applied *Survival* Analysis Using R. Springer. USA

Prihantoro, Imam Basuki, Kasir Iskandar. (2013) Analisis Faktor-Faktor Makro Ekonomi dan Demografi Terhadap Fungsi Permintaan Asuransi Jiwa di Indonesia. Jakarta Volume 1 Nomor 1

OJK. (2017). Statistik Asuransi - Desember 2017 https://www.ojk.go.id/ id/kanal/iknb/data-dan-statistik/asuransi/Pages/Statistik-Asuransi---Desember-2017.aspx

OJK. (2017). Statistik Perasuransian Indonesia - 2016 - Revised Version https://www.ojk.go.id/id/kanal/iknb/data-dan-statistik/asuransi/Pages/Statistik-Perasuransian-Indonesia---2016---Revised-Version.aspx

Republik Indonesia. (2014). Undang-Undang No. 40 Tahun 2014 tentang Perasuransian

Rodriguez German. (2005). Non Parametrik Estimation in *Survival* Models

Schneider C Judith, Sven Nolte.(2017).Don't *Lapse* Temptation: A behavioral explanationfor policy Surrender. Journal of Banking and Finance.Germany

Shacham Mordechai, Brauner Neima. (2014) Application of stepwise regression for dynamic parameter estimation.Computers and Chemical Engineering. Israel

Tableman, Mara; Kim, Jong Sung. Survival Analysis Using S: Analysis of time-to-event data. 2004. Chapman & Hall. Boca Raton