Int. J. Nonlinear Anal. Appl. 12 (2021) No. 1, 1243-1250
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5003

Application of the accelerated failure time model to lung cancer data

Akam Ali Othman[a], Sabah Haseeb Hasan[a,*]

[a] *College of Administration and Economics\University of Kirkuk*

Abstract

Accelerated failure time model sometimes symbolized as AFT model, is an important regression model in survival analysis. In this article, we applied AFT model to the data of lung cancer patient in order to identify the must important factors affecting the patient's survival time. The results showed a well performance for this model, as based on some statistical criteria, the factors that are consistent with the opinion of specialists in influencing survival time were identified, as the factors (smoking, treatment, proliferation, location of residence) of the main factors affecting the life of a person with this disease.

Keywords: Accelerated failure time model, life time, survival data, selection criteria, lung cancer.

1. Introduction

Accelerated failure time model sometimes symbolized as AFT model, is an important regression model in survival analysis (Khanal et al., 2014). This model is sometimes applied in reliability analysis in industrial experiments, it is used as an alternative to the Cox regressionmodel in the medical field where a better description and interpretation is obtained (Yamaguchi,1992). Using this model, the explanatory variables that have an effect on the acceleration or deceleration of the time to hold until the occurrence of an event are determined (Pan, 2001) This model is considered one of the statistical techniques that can deal with censoring data, The accelerated failure time model is also characterized by an attention to the time that affects patient survival. However, it is accelerating or slowing, assuming that the dependent variable is binary. The model is used in the following cases (Wei, 1992).

1. It uses an accelerated failure time model when the dependent variable is two-response and Interested in staying time.

*Corresponding author

Email addresses: akam.ali@uokirkuk.edu.iq (Akam Ali Othman), sabahsaqi@uokirkuk.edu.iq (Sabah Haseeb Hasan)

Received: January 2021 *Revised:* March 2021

2. It uses an accelerated failure time model to predict survival time and determine an individual's risk level.
3. It is used to study the effects of explanatory variables that influence survival time.
4. It is used for a comparative study between two or more types of treatment for a specific disease.

This article amins to apply the AFT model to analyze lung cancer patient's data and testing the validity of the estimated model in identifying the important factors that affect the patient survival time.

2. Accelerated Failure Time (AFT) Model

The formula for the AFT model is as follows (Kay & Kinnersley, 2002)

$$T = e^{\mu} + e^{\hat{\beta} x_i} + e^{\sigma \epsilon_i} \quad i = 1, 2, 3, \ldots, p \tag{2.1}$$

Where the risk function of the accelerated failure time model is as follows:

$$h_i(t) = e^{-\eta i} h_0 \left(\frac{t}{e^{\eta i}} \right) \tag{2.2}$$

Where $\eta_i = \hat{\beta} x_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$
h_0 It is the primary risk function.
The survival function of the accelerated failure time model is as follows:

$$\begin{aligned} S_i &= P\left(T_i \geq t\right) \\ &= P\left\{ \exp\left(\mu + \dot{\beta} x_i + \sigma \epsilon_i \right) \geq t \right\} \end{aligned} \tag{2.3}$$

The survival function can be written as following:

$$\begin{aligned} S_i &= P\left(T_i \geq t\right) \\ &= P\left\{ \exp\left(\mu + \sigma \epsilon_i\right) \geq t / \exp\left(\hat{\beta} x_i\right) \right\} \end{aligned} \tag{2.4}$$

$$S_i(\mathrm{t}) = S_0 \left\{ \frac{t}{\exp\left(\eta_i\right)} \right\} \tag{2.5}$$

S_0 is the primary survival function depends on time.
It is also possible to use transfers for acceleration failure time model it gives great clarification and facilitation, to understand and interpret the model and its effect for explanatory variables on the time preceding the occurrence of the event. Which affects either accelerating or slowing down (Orbe et al., 2002), where the function is converted by taking the logarithm to it and in the following (Walker & Mallick,1999).

$$Y = \log(T) = \mu + \dot{\beta}_i x_i + \sigma \epsilon_i \tag{2.6}$$

Where Y is the two-response dependent variable; μ is constant paramiter; β is the vector form parameter; x_i is the explanatory variables; $\sigma \epsilon_i$ is the error with a fixed limit.

Lawless (1982) estimated the parameters of the accelerated failure time (AFT) model. Which determines the relationship between the explanatory variables available for the studied item (Daviad, 2015), And compute the risk function and the survival function, The parameters of the AFT model are estimated in maximum likelihood method as follows (Kestenbaum, 2019)(Wei, 1992)

$$L(\beta, \mu, \sigma) = \prod_{i=1}^{n} \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i} \tag{2.7}$$

Where $f_i(t_i)$ and $S_i(t_i)$ are the survival probability and density functions of an individual (i) in time $(t_i), \delta_i$ is the event index for the individual (i), Where the event indicator is equal to the correct one when the event (death) occurs and zero if the individual is subject to censoring (Pan, 2001).

$$S_i(t_i) = S_{\epsilon i}(z_i)$$

Where $z_i = (\log t_i - \mu - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi}) / \sigma$

$$f_i(t_i) = \frac{1}{\sigma t_i} f_{\epsilon i}(z_i)$$

Thus, it is possible to write the maximum likelihood function of survival probability as follows:

$$L(\beta, \mu, \sigma) = \prod_{i=1}^{n} (\sigma t_i)^{-\delta_i} \{f_{\epsilon i}(z_i)\}^{\delta_i} \{S_{\epsilon i}(z_i)\}^{1-\delta_i} \tag{2.8}$$

Then you take the natural logarithm of the maximum likelihood function as follows:

$$\log L(\beta, \mu, \sigma) = \prod_{i=1}^{n} \{-\delta_i \log(\sigma t_i) + \delta_i \log f_{\epsilon i}(z_i) + (1-\delta_i) \log S_{\epsilon i}(z_i)\} \tag{2.9}$$

From the above formula the logarithm of a maximum likelihood function is derived to (P + 2) from the first times the derivative, and after obtaining the product of these derivations (Huang et al., 2006), it is equal to zero (P + 2) is obtained from the equations, by using the iterative method (Newton Raphson) to solve the equations, we obtain estimates of the parameters of the accelerated failure time model (Zeng & Lin, 2007).

3. Model Evaluation

The first step in the process of evaluating the fit of the model that has been fitted is usually the evaluation of the significance of the model, that is, the effect of the explanatory variables is determined as a whole in the model (Faruk, 2018). This will be taken in the following sections.

3.1. Likelihood Ratio Test

Likelihood ratio (L.R.) is a test through which a decision is made about the relevance of the model as a whole. It also has importance in knowing which of the models is more suitable for the studied data, this test is as follows

$$LR = -2\log\left(\frac{L_M}{L_0}\right) = 2\log L_0 - 2\log L_M$$

Where L_0 is the log likelihood when only the primary risk function is present in the model, L_M is log likelihood when there are (M) of the explanatory variables with in the model.

The value of the test is compared to a chi-square distribution with a degree of freedom equal to the number of variables in the model (Wienke, 2010).

3.2. Model selection Criteria

Akaike information criterion (AIC) is considered one of the criteria for selecting the best model, this is done by calculating the criterion value for the models [3].The model with the lowest value for the criterion is the best.This criterion is defined as follows (Akaike, 1974)

$$AIC = -2 \log L_M + 2k$$

$\mathrm{L_M}$ is log likelihood when there are (M) of the explanatory variables within the model, k is the number of explanatory variables included in the model.
Another criterion can be used to select best model is the Baysian information criterion (BIC), this criterion can be calculated as follow (Schwarz, 1978)

$$BIC = -2 \log L_M + (k) \log(n)$$

Where n is the ample size. The model with the least value is the best.

3.3. Test Significance of Variables

To determent the significance of the explanatory variables, this dependants on Wald Test. Where after choosing the model as a whole the parameters of the estimated model have to be tested (Cleves et al., 2008), The Wald test is one of the methods used to test the parameters accompany explanatory variables that entered within the model. The null hypothesis for this test states that the parameter is equal to zero, as following (Faruk, 2018)

$$H_0 : b_j = 0$$
$$H_1 : b_j \neq 0$$

Wald Test can be calculated as following:

$$W^2 = \left(\frac{\hat{b}_j}{s \cdot e_{\hat{b}_j}} \right)^2$$

Where $\hat{b}_j$ is the estimated value of a parameter of the explanatory variable $\mathrm{X_j}$
S. $e_{\hat{b}_j}$ is the standard error of a parameter of the explanatory variable $\mathrm{X_j}$.
As the value of Wald test follows the chi-square distribution with one degree of freedom.

4. Application to Lung Cancer Data

In this article, the sample represents patients with lung cancer in Kirkuk governorate-Iraq. A random sample was selected for patients with the disease from Azadi teaching hospital and the specialized center for oncology and hematology. As the information was taken from the files of (103) patient's with lung cancer, it included the time period from (1/1/2017) to (1/9/2020), The sample also included the number of cases of censored (43) persons and the number of deaths (60) persons. The information was recorded about the time of the disease surviving since the disease, the age of the patient, the gender of the patient, the cessation of smoking, the type of treatment that the patient receives, the proliferation of the spread of the disease and finally the patient's residence in the city center or in the city parties. After collecting information on patient's with lung cancer, the variables that will be used in estimating the (AFT) model were defined as follows
T: Survival time, which is the time of survival of patients with lung cancer until the occurrence of

an event (death) or cencoring.
Y_i : (Consoring); i.e. the final condition of the patient:(0) Consored;(1): (death).
X_1 : Age of the patient.
X_2 : Gender;(1) Male,(2) female.
X_3 : Smoking; (0) non-smoker;(1) Smoker.
X_4 : Treatment;(1) Chemotherapy,(2) Radiotherapy.
X_5 : Proliferation; (0) Not diffuse, (1) Diffuse.
X_6 : location; (1) City center, (2) City parties.

It is worth noting that a program has been designed to calculate the indicators of the AFT model using the matlab program.
Before strarting to estimate the parameters of (AFT) model, it is necessary to know distribution that the error follows. As the error often follows the distributions (Exponential, Log-normal,Weibull, Log-logistic). The determination of the distribution of the error is based on the (AIC) and (BIC) criteria, as show in tabel 1 .

5. Results and Discussion

Table 1 shows the criteria and testing for the distribution of error, through the results of the table 1 , by comparing the values of AIC and BIC, it becomes clear that the error follows the log-logistic distribution, due to the possession of the lowest value for the AIC criteria, which equals 179.6858 , and the lowest value for the BIC criteria, which equals 200.7636 . Thus, the distribution of the error is the Log-logistic.

Table 1: Determination the distribution of Error for (AFT) Model.

Model	No. of parameters (p)	$\log L_M$	AIC	BIC
Exponential	7	-101.0792	216.1583	234.6014
Weibull	8	-89.9003	195.8005	216.8784
Log-logistic	8	-81.8429	179.6858	200.7636
log-normal	8	-82.2267	180.4534	201.5313

It is noted from the results in table 2 . where using the backward Wald method that analyzes the variables and then determines the largest non-significant variable and excludes it from the model to form a new model, the program will repeat the same process until the remaining variables are all significant, then the analysis stops, so that the remaining variables in the model are all significant variables according to the type of study and the type of explanatory variables in the analysis.
Also through table 2 and by comparing the results of the column Wald test with the tabular value, which is distributed chi-square with one degree of freedom and the level of significance (0.01) , it was found that the gender variable in the first step was one of the largest non-significant variables through comparison with the significant level or through the Wald test. therefore, it is excluded from the model and the analysis is done again with the presence of five explanatory variables (age, smoking, treatment, proliferation, location) in edition the constant term.
As for the step-2 of the analysis, the (AFT) model, we notice that the age variable is of the largest non-significant level in the new model, so it is excluded and the analysis is done again with the presence of four explanatory variables (smoking, treatment, proliferation, location).
After excluding the gender variable in the step-1 and excluding the age variable in the step-2, all the remaining variables in the model are significant variables, so the final model contains the following

explanatory variables (smoking, treatment, proliferation, location) in edition the constant term. Returning to the table 2 , we note that there are three models for the accelerated failure time (AFT), as it is important to determine which of the three models is the largest significant model among the models, and this is done by calculating the values of the likelihood retio test, a comparison with chi-square distribution with a degree of freedom equal to the number of parameters model as shown in table 3 . From Table 3, specifically from the results of the likelihood ratio test, and in comparison with the tabular values, the null hypothesis of the three models was rejected, which states that all parameters of the explanatory variables in the model have no effect on the dependent variable, so the alternative hypothesis has been accepted which states that the parameters of the explanatory variables entering the model have a significant effect on the dependent variable. As for which model is the best, the model in the Step- 3 is the best.Also by referring to the results of table 2 through the Wald test, the same model in step- 3 is the best.

Table 2: Results of data analysis and statistical indicators of the accelerated failure time model according to backward wald method

Steps	Variable's in model	β	S.E (β)	Wald	d.f	Sig.
Step-1	Constant	5.4612	0.7967	46.9844	1	0.000
	Age	-0.0095	0.007	1.8396	1	0.175
	Gender	-0.1718	0.2253	0.5819	1	0.445
	Smoking	-0.6989	0.2375	8.6630	1	0.003
	Treatment	0.4821	0.2072	5.4126	1	0.02
	Proliferation	-0.6635	0.1968	11.367	1	0.0007
	Location	0.6124	0.1535	15.9153	1	0.0001
Step-2	Constant	5.1579	0.6863	56.4848	1	0.000
	Gender	-0.0099	0.007	2.0195	1	0.1553
	Smoking	-0.5625	0.1509	13.8954	1	0.0002
	Treatment	0.4976	0.203	6.0095	1	0.0142
	Proliferation	-0.665	0.1932	11.8438	1	0.0006
	Location	0.5974	0.1510	15.6489	1	0.0001
Step-3	Constant	4.3808	0.4363	100.7940	1	0.0000
	Smoking	-0.5993	0.1517	15.6093	1	0.0001
	Treatment	0.5802	0.201	8.3271	1	0.0039
	Proliferation	-0.6803	0.1994	11.638	1	0.0006
	Location	0.6383	0.152	17.6445	1	0.0000

Table 3: Statistical indicators for selection the best AFT model

Steps	$\log L_0$	$\log L_M$	LR test	D.f.	χ^2_{tabel}	AIC	BIC
Step-1	-120.9616	-81.8429	78.2374	6	16.81	175.69	175.76
Step-2	-120.9616	-82.1428	77.6377	5	15.08	172.29	174.35
Step-3	-120.9616	-83.1346	75.6541	4	13.28	170.27	174.32

With regard to the (AIC) and (BIC) criterion, the values of each of the AIC criterion and the BIC criterion for the accelerated failure time models are evident. That is the model in the step-3 is the best on each of (smoking, treatment, prolieration and location) as significant explanatory variables. As for the accelerated failure time (AFT) model, the value of the AIC criterion was equal to (170.27) for the accelerated failure time model, while the BIC criterion was equal to (174.32). Thus this model is best model according to statistical indicators which includes four explanatory variables (Smoking, treatment, proliferation, and location). Accordingly the mathematical formula for the accelerated failure time model can be written as follows:

$$\log(T) = 4.3808 - 0.5933X_1 + 0.5802X_2 - 0.6803X_3 + 0.6383X_4$$

With regard to the relevance of the statistical results to the medical reality, the accelerated failure time (AFT) model kept the proliferation variable in the model. According to the opinion of the specialists in this disease, this variable is considered essential in diagnosing the lung cancer and influencing the time of the patient's survival as well as other variables in the same model. In addition, it is noted that the prolieration variable had a negative effect in the accelerated failure time model,this is evidence of the importance of using this model in representation the data of lung cancer, so was the case with the smoking variable.

6. Conclusions

The article dealt with applying the accelerated failure time model to the data of lung cancer patients in order to identify the most important factors affecting the patient's survival time. We reached the following conclusions:

1. Through the results of applying the accelerated failure time model, it was found that each variable (smoking and proliferation) had a negative effect on the survival time,this results is in agreement with the medical opinion.
2. Also through the results of applying the accelerated failure time model, it was found that each variable (treatment and location) have a positive effect on the survival time, this results is in agreement with the medical opinion.
3. In general, the accelerated failure time model has well performance in analyzing and identifying the most logical factors affecting the final state of a patient with lung cancer.

References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE. Trans. Aut. Cont. 19(6) (1974) 716–723.

[2] M. Cleves, W. W. Gould, R. Gutierrez and Y. Marchenko , *An introduction to survival analysis using stata*, Stata Press, 2008.

[3] C. Daviad, *Modelling Survival Data in medical research*, CRC Press, Third Edition, New York, 2015.

[4] A. Faruk, *The comparison of poportional hazards and accelerated failure time models in analysis the first birth interval survival data*, J. Phys. Conference Series, 974(1), IOP Pub. 2018.

[5] J. Huang, S. Ma and H. Xie, *Regulaized estimation in the accelerated failure time model with high-dimensional covariates*, Biomet. 62(3) (2006) 813–820.

[6] R. Kay and N. Kinnersley, *On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data a case study in influenza*, Drud Inf. J. 36(3) (2002) 571–579.

[7] B. Kestenbaum, *Epidemiology and biostatistics*, An introduction to Clinical Research, Second Edition, 2019.

[8] S. P. Khanal, V. Sreenivas and S. K. Acharya, *Accelerated failure time models: an application in the survival of acute live failuer patients in India*, Int. J. Sci. Res. 3 (2014) 161–66.

[9] J. Orbe, E. Ferreira and V. Nunez-Anton, *Comparing proportional hazards and accelerated failure time models foe survival analysis*, Stat. Med. 21(22) (2002) 3493–3510.

[10] W. Pan, *Using frailties in accelerated failuer time model*, Lifetime Data Anal. 7(1) (2001) 55–64.

[11] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. 6 (1978) 461–464.

[12] S. Walker and B. K. Mallick, *A Bayesian semiparametric accelerated failure time model*, Biomet. 55(2) (1999) 477–483.

[13] L. J. Wei, *The accelerated failure time model: a useful alternative to the cox regerssion model in survival analysis*, Stat. Med. 11(14-15) (1992) 1871–1879.

[14] A. Wienke, *Frailty models in survival analysis*, CRC press.(2010).

[15] K. Yamaguchi, *Accelerated failure-time regression models analysis of permanent employment in Japan,* J. Amer. Stat. Assoc. 87(418) (1992) 284–292.

[16] D. Zeng and D. Y. Lin, *Efficient estimation for the accelerated failure time model*, J. Amer. Stat. Assoc. 102(480)(2007) 1387–1396.