ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202204_5(4).0032

Variable Selection Based on Kaplan-Meier Estimation and Accelerated Failure Time Model

Yanzhao Zeng[1, *]

[1]Department of statistics, School of Economics, Jinan University, Guangzhou, 510632, China

Abstract

Variable selection is to use statistical methods to select the most suitable subset from a large number of variables to explain the model and predict. Identifying and selecting appropriate variables is more important in model prediction. Based on Kaplan-Meier (KM) estimation and accelerated failure time (AFT) model, this paper explores whether 'age at surgery'and 'medication' have an impact on the postoperative survival of breast cancer patients, and then discusses the impact of variable selection on actual problems. The analysis found that the effect of actual age on survival time needs to be considered, so time-dependent covariates are used for improvement. On the basis of the existing covariates, a variable is added to represent the actual age as the time-dependent covariate of the model. The results show that the time-dependent covariates make up for the deficiencies in some aspects, making the analysis more comprehensive and reasonable.

Keywords

Kaplan-Meier estimation; Variable selection; Accelerated failure time model.

1. Introduction

Survival analysis models widely appear in the fields of medicine, biology, insurance and other scientific fields, and mainly study the relationship between the time of event occurrence and certain events. This type of model has two characteristics: one is with censored data, that is, because some observed individuals drop out or cannot be tracked, the survival time of some individuals cannot be observed, so there are censored data. The second is high-dimensional data. Due to the large amount of data generated in bioinformatics, medicine and other fields and the rapid development of computer technology, there is inevitably a large amount of survival data in this part of the field. For individuals, the factors that affect lifespan time There are also a large number of latent variables. For example, in the colon cancer study [1], in order to compare the effect of adjuvant drug levamisole treatment with levamisole and fluorouracil mixed treatment after colon cancer resection, the time of cancer recurrence was different for each patient according to the physical condition of each patient. Various physical indicators such as height, weight, medical history, and blood type may be relevant explanatory variables.

Variable selection is the use of statistical methods to select the most appropriate subset of variables to explain the model and predict. It has three main functions: one is to eliminate the relevant and redundant variables in the model. The existence of these variables cannot improve the accuracy of the model, so it is necessary to do relevant processing; Thereby, the accuracy of the model is improved [2]; the third is to reduce the dimension of the model and the unnecessary calculation amount. With the increase of the sample size and dimension, the calculation amount of the model, especially the nonlinear model will increase exponentially, and the variable selection will greatly reduce the computational cost.

To analyze survival data, statisticians prefer to choose Cox proportional hazards regression models or accelerated failure time models, which are considered the most popular models in survival analysis. The accelerated failure time model assumes that the failure time has a linear relationship with the covariates after logarithmic transformation, because its model structure and interpretation of regression parameters are similar to general linear regression equations [3]. Compared with the Cox proportional hazards regression model, the interpretation of the results is also simpler, more intuitive, and easier to accept. Because of the advantages of the accelerated failure time model, many statisticians and scholars have done a lot of research on the estimation of unknown parameters in the model, the prediction of the model and the applicable conditions. So far, scholars have done a lot of research on accelerated failure time models under censored data, including right-censoring and interval-censoring, and they have also achieved fruitful academic results.

The accelerated failure time model can relate the logarithmic form of the failure time and the covariates in the form of a linear relationship. Compared with the Cox model and the additive hazard model, the representation is simpler and more straightforward. When the covariate increases or decreases, the failure time T speeds up or slows down. Regarding the accelerated failure time model, many scholars have given relevant research. Such as: Pan (2001), Lambert et al. (2004) and so on. Pan (2001) gave Frialty's method to describe the possible correlation and difference of failure time for multivariate failure time in the study of accelerated failure time model, and used the similar algorithm of EM to estimate the parameters of Frailty model. However, the authors did not study the relationship between observation time and failure time. Lambert et al. (2004) used an accelerated time-to-failure model to study 31 kidney transplant patients in the United Kingdom and identify prognostic factors. The model can combine different explanatory variables and random effects into one, and the model has a good result for transplanted data.

Based on Kaplan-Meier (KM) estimation and accelerated failure time (AFT) model, this paper proposes variable selection, which is verified with actual data, and then discusses the impact of variable selection on practical problems.

2. Methodology

2.1. Kaplan-Meier Estimation

Assuming that in a group of observation objects, by the end of the observation, a total of *m* individuals died at *j* time points (or any other outcome event under study, hereinafter, death is used as an example), and the order of death time points is: $0\leq t(1) \leq t(2) \ldots\leq t(j) \leq \infty$. There is a $n=n_0$ sample from the survival function *S* to be estimated. It is assumed that just before a certain death time *t(j)*, there are still r_j observation objects that are at risk of death (meaning that they are still alive and have not been censored), and d_j death occurs at the time *t(j)*, Based on this, the survival function value at time *t* can be estimated as:

$$\hat{S}(t)=\prod_{j:\, t_j\leq t}\left(1-\frac{d_j}{r_j}\right) \tag{1}$$

This value is also called the K-M estimator. It is not difficult to see that the K-M estimate should be a discrete function in the time dimension by definition. If it is assumed that all death events occur exactly at the time of death, and no death occurs between two discrete death time points, the survival curve can be drawn as a continuous, gradually decreasing step function according to the KM estimate, Only changes the value at the point of death. For discrete time points, it can be shown that the K-M estimate is actually the maximum likelihood estimate [4].

The K-M estimate can be shown to follow an approximately normal distribution [5], so its confidence interval can be calculated. Greenwood proposed an approximate formula for calculating its confidence interval (Greenwood's formula):

$$\widehat{V}\left(\hat{S}(t)\right) \approx \left[\hat{S}(t)\right]^2 \sum_{j:t_j \le t} \frac{d_j}{r_j(r_j - d_j)} \tag{2}$$

This formula can also be used to calculate percentile survival time confidence intervals.

2.2. Accelerated Failure Time Model

Let T be the time to failure and X be the corresponding covariates. Without censoring observations, we can directly study the regression equation of T with respect to X :

$$T_i = \mu + x_i\beta + \varepsilon_i \qquad i=1, 2, \ldots, n$$

where μ is the constant term, ε_i is the error term, and $\beta$is the regression coefficient of the covariate.

However, censored values are often present, and it was later found that if a log-transformation of survival time T is used as the dependent variable, it can be used to analyze cases with censored data. Let be the survival time without covariates, its hazard function form is:

$$h_0(t) = \frac{f_0(t)}{1-F_0(t)} \tag{3}$$

Now assume that with covariate X, the individual survival time $T=e^{x^T\beta}T_0$, in this relationship, if $x^T\beta<0$, then T is smaller than T_0, that is, the covariate accelerates the individual's failure process. The hazard function has the form

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{f_0(e^{-x^T\beta}t)e^{-x^T\beta}}{1-F_0(e^{-x^T\beta}t)} = h_0(e^{-x^T\beta}t)e^{-x^T\beta} = \phi h_0(\phi t) \tag{4}$$

ϕ is called the acceleration factor. Apply a logarithmic transformation to both sides,

$$\log(T) = x^T\beta + \log T_0 = x^T\beta + \varepsilon \tag{5}$$

The above formula is the accelerated failure time model.

The model structure of the accelerated failure time model and the interpretation of the regression coefficients are similar to the general linear regression equation, and the analysis results are easy to understand, and the error term does not specify a specific distribution form, which shows its practicability and flexibility.

When applying the parametric method, people usually express the AFT model as:

$$Y_i = \log T_i = \mu + x^T\beta + \sigma\varepsilon_i \tag{6}$$

where β is the regression coefficient vector and ε is the random error. In practice, it has been found that there are four models that can be attributed to the AFT model family, namely Weibull regression model, logarithmic logistic regression model, logarithmic normal distribution regression model and generalized Gamma distribution regression model.

Taking the Weibull regression model as an example, when the error term of the accelerated failure time model is assumed to obey the Weibull distribution, the probability density function of ε_i=(log T_i= $\mu+x^T\beta)/\sigma$ is:

$$f_\varepsilon(\varepsilon_i)=\exp\{\varepsilon_i-\exp(\varepsilon_i)\}$$

the probability density function of Y_i= log T_i= $\mu+x^T\beta+\sigma\varepsilon_i$ is:

$$\frac{1}{\sigma}f_\varepsilon\left(\frac{\log x\text{-}\ \mu\text{-}x^T\beta}{\sigma}\right)$$

Then, the probability density function of the survival time of the i-th observation object is:

$$f_i(y)=\frac{1}{\sigma}\ \exp(u_i\text{-}e^{u_i})$$

where $u_i=\frac{y\text{-}\mu\text{-}x^T\beta}{\sigma}$

The likelihood function of the log survival function based on n observation units is:

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

f_i and S_i are the probability density function and survival function of the i-th object at log(t_i), respectively,

δ_i indicates the survival state, 1 indicates failure, and 0 indicates censoring.

3. Modelling

3.1. Survival Function Based on K-M Estimation

The data in this section are derived from the data set of Reference 7, and the purpose is to explore whether 'age at surgery' and 'medication' have an impact on the survival of breast cancer patients after surgery. The data contains 4 variables: time, index, age and group, which represent survival time (unit: year), censoring (1 means death, 0 means censoring), age at surgery and medication or not (2 means use anticancer drugs, 1 means no anticancer drugs).

For the given data, we can calculate the survival function through the above formula (2.1), and then draw the survival function curve. In order to further compare the differences of survival functions more precisely, the confidence interval of each point can be calculated, and the confidence interval of the two groups of samples at the same time point can be compared. First, the calculated variance of $\hat{S}(t)$ is

$$\hat{V}_S(t)=\widehat{Var}[\hat{S}(t)]=\hat{S}^2(t)\sum_{j:t_j\leq t}\frac{d_j}{n_j(n_j\text{-}d_j)}$$

The 95% confidence interval for $\hat{S}(t)$ is

$$[\hat{S}(t)]^{\exp[\pm1.96\hat{S}(t)]}$$

3.2. Modeling with Accelerated Failure Time Model

As mentioned above, the AFT model of the parametric method includes Weibull , exponential distribution, lognormal and logarithmic logistic model according to different distributions. Using these four models, the estimated values of the coefficients of the two covariates age and group and the log-likelihood of each model were obtained. After the Log-Rank test, the coefficients of the covariates were finally analyzed to obtain their influence on the survival time. In addition, the analysis found that the effect of actual age on survival time needs to be considered, so time-dependent covariates were used for improvement. On the basis of the existing covariates, a variable representing actual age was added as a time-dependent covariate of the model.

4. Numerical Study

4.1. Visualized Survival Curves Obtained by K-M Estimation

For the treatment variable, plot the survival function curve and confidence region of the two groups of samples.

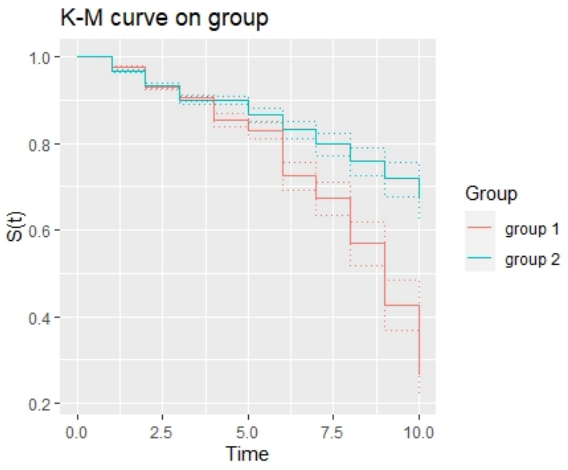

Figure 1. KM curve based on group variable

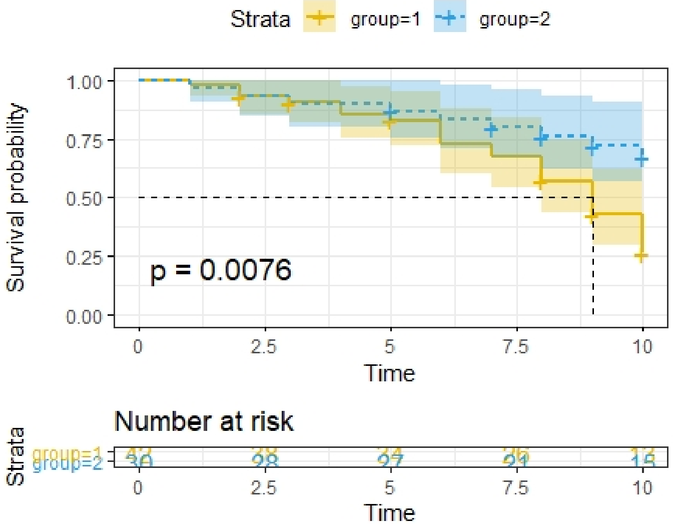

Figure 2. KM curve based on group variable, p-value=0.0076.

The above results are obtained by KM estimation. Combined with Figure. 1 and Figure. 2, the results show that the survival function curve of group2 is above the survival function curve of group1, and the p-value is 0.0076, which is less than 0.05, indicating that under the test level of 0.05, The KM estimation results are significant, and there are significant differences in the survival curves between groups, that is, group has a significant impact on the survival status of breast cancer patients.

4.2. Variable Selection Based on AFT Model

4.2.1. Covariates under Four AFT Models

Table 1. Covariate coefficients under the weibull distribution

-	Value	Std. Error	z	p
(Intercept)	5.1354	0.5884	8.73	< 2e-16
age	-0.0668	0.0111	-6.02	1.7e-09
group	0.4161	0.1330	3.13	0.0018
Log(scale)	-1.1161	0.1346	-8.29	< 2e-16

Table 2. Covariate coefficients under exponential distribution

-	Value	Std. Error	z	p
(Intercept)	7.2849	1.2789	5.70	1.2e-08
age	-0.1171	0.0222	-5.28	1.3e-07
group	1.0024	0.3852	2.60	0.0093

Table 3. Covariate coefficients under lognormal distribution

-	Value	Std. Error	z	p
(Intercept)	5.35067	0.45458	11.77	< 2e-16
age	-0.07481	0.00867	-8.63	< 2e-16
group	0.44346	0.14613	3.03	0.0024
Log(scale)	-0.80364	0.11804	-6.81	9.9e-12

Table 4. Loglogistic accelerated failure time model

-	Value	Std. Error	z	p
(Intercept)	5.15640	0.49793	10.37	< 2e-16
age	-0.07167	0.00959	-87.47	7.8e-14
group	0.46377	0.13993	3.31	0.00092
Log(scale)	-1.39884	0.13533	-10.34	< 2e-16

The coefficients and log-likelihoods of the covariates fitted by the four models are obtained through the above, and the results are shown in the following table:

Table 5. The log-likelihood corresponding to the different models

AFT	Value	df	p
Weibull	-94.4	2	7.6e-14
exponential	-116.6	2	1.8e-08
lognormal	-95.3	2	3e-15
loglogistic	94.8	2	1.6e-14

The coefficients obtained for each model were different but with the same sign, with a negative coefficient for age and a positive coefficient for treatment. Changing the speed of time by this coefficient, the time axis is stretched when $\exp(x^T\beta)>1$; when $\exp(x^T\beta)<1$, the time axis is compressed. For the treatment mode, the covariate coefficient is positive, so the acceleration factor ϕ <1 and the acceleration factor of group1 is greater than that of group2, which is reflected in the survival function image that the survival function curve of group1 is below the survival function curve of group2.

The P values of the four models are all less than 0.05, indicating that the models are significant at the test level of 0.05. The P values corresponding to the covariate coefficients are all less than 0.05, indicating that if the significance level is 0.05, both age and group have significant effects on the survival of breast cancer patients. However, it is not difficult to find that the actual age of the patient changes with time. If only the age at the time of surgery is considered and the influence of the actual age on the risk rate is ignored, the results will be insufficient. In order to make the results more comprehensive, it is necessary to consider the effect of chronological age on the risk rate, and to improve the AFT model. On the basis of the existing covariates, a variable representing chronological age should be added as a time-dependent covariate of the model.

4.2.2. Newage under Four AFT Models

Table 6. The log-likelihood corresponding to the different models

AFT	Newage/value	Value	p
Weibull	-0.0564	-112.4	7.6e-07
exponential	-0.1107	-125.3	1.9e-05
lognormal	-0.0775	-112.7	1.5e-08
loglogistic	-0.0679	-112.7	1.4e-07

Table 7. Log-Rank test

	Chisq	df	p
age+group	192	48	<2e-16
Newage	108	26	6e-12

The p-value for chronological age (newage) was less than 0.05, indicating that chronological age did have a significant effect on the hazard rate, while the coefficient for chronological age (newage) was negative, indicating an acceleration factor. Therefore, in actual analysis, such data should not be ignored, but the actual age (newage) should be taken into account while considering the influence of age at the time of surgery and the effect of medication use (group) on the survival of breast cancer patients. The effect of this time-dependent covariate on survival in breast cancer patients. It also shows that the time-dependent covariates make up for the deficiencies in some aspects, making the analysis more comprehensive and reasonable.

5. Conclusion

By analyzing the difference and relationship between time-dependent covariates and original independent covariates, variable selection was studied based on R software. From the analysis results, when there are variables that change with time in the data, the time-dependent covariate AFT model can make up for the deficiencies in some aspects, making the modeling more reasonable and the model analysis more comprehensive.

Acknowledgments

This authors would like to thank the editor, the referees for their constructive comments.

References

[1] C.G. Moertel, T.R.Fleming, J.S. Macdonald, et al.(1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. New England Journal of Medicine, 322(6), p.352-358.

[2] D. Liu (2016). Research on variable selection in Cox model and variable coefficient Cox model. Jinan University.

[3] X.H. Yuan and J. Chen (2017). Empirical Likelihood-Based Weighted Estima-tion of Accelerated Failure Time Models with Missing Covariates. Journal of Northeast Normal University(Natural Science Edition), (4), p.32-37.

[4] DR Cox and D Oakes(2018). Analysis of survival data. Chapman and Hall/CRC.

[5] Y.Y. Xiao, C.Z. Xu and N.Q. Zhao (2016). Commonly used survival analysis models and their estimation methods for time-dependent covariate effects. Chinese Journal of Health Statistics, 33(3), p.543-547.

[6] Z. Jin, D.Y. Lin, L.J. Wei, et al. (2003) Rank-based inference for the accelerated failure time model. Biometrika, 90(2), p.341-353.

[7] J.P. Li and B.C. Xie (2008) Multivariate Statistical Analysis Methods and Applications. China Renmin University Press, Beijing.

[8] Pan, W(2001). Using frailties in the accelerated failure time model. Lifetime Data Analysis, 7, p.55-64.

[9] Lambert, P., Collett, D., Kimber, A., and R. Johnson, R. (2004), Parametric accelerated failure time models with random effects and an application to kidney transplant survival, Statistics in Medicine, vol. 23, p.3177-3192.