Efficient Algorithms for Computing the Impact of Factors on Survival Time and the Confidence Intervals

Daxin Zhu[1], Xiaodong Wang*[2], and Jun Tian*[3]

[1]Faculty of Mathematics & Computer Science, Quanzhou Normal University, China
[2]Fujian University of Technology, Fuzhou, China
[3]School of Public Health, Fujian Medical University, Fuzhou, China
[1]dex@qztc.edu.cn; [2]wangxd@qztc.edu.cn

***Abstract*- In some researches, time is a target variable. Factors that may influence the occurring time of an outcome need to be analysed. The effect of a factor on an outcome is often modified by another factor because there is an interaction between them. The analysis of the interaction between the factors is very important for us to better understand the mechanism of the effect that factors exert on an outcome. This paper proposes the method to evaluate interactions of the factors and their 95% confidence intervals in survival analysis. These factors are influencing the survival time of patients with cancer, and their interactions are successfully analysed by the method.**

Keywords-*Interaction; Survival Time; Confidence Intervals; Survival Analysis*

I.INTRODUCTION

Survival time is main observation target in the researches on life issues [6]. The connotation of survival time is in general. It could be a survival time of patients or animals, and it could be life of the product, or the time of something changed from state to state. No matter what the field of research, the corresponding data analysis methods are referred to as survival analysis when the time variable is a main observation target. Survival analysis is widely used in many research areas such as biomedical, industrial, agriculture, forestry and other industries. In many cases, survival time is affected by many factors. For example, the survival time of cancer patients can be affected not only by the treatment, but also by the clinical stage, pathological type, nutrition, mental status, and many other factors. It is very important to identify the influence factors of survival time for improving the survival time and patient outcomes.

Let $x_1, x_2, \cdots, x_m$ be m factors affecting the survival time. The observation target is the time when an event A occurs in a follow-up study on the objects. If the risk of the occurrence of event A at time t is denoted as $h(t, x)$, then a model of the relationship between factors and the risk of the occurrence of event A can be described as

$$h(t, x) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m} \tag{1}$$

This formula was proposed by Cox [2], where $h_0(t)$ is called a base risk, which is the basic incidence rate if $x_1, x_2, \cdots, x_m$ have no effect on survival time. There are m parameters $\beta_1, \beta_2, \cdots, \beta_m$ in the model. The formula (1) is also called proportional hazards model.

If the m factors $x_1, x_2, \cdots, x_m$ take the values $a_1, a_2, \cdots, a_m$, then the risk of the occurrence of event A of the observed objects must be $h(t, a) = h_0(t)e^{\beta_1 a_1 + \beta_2 a_2 + \cdots + \beta_m a_m}$. On the other hand, if the m factors $x_1, x_2, \cdots, x_m$ take the values $b_1, b_2, \cdots, b_m$, then the risk of the occurrence of event A of the observed objects must be $h(t, b) = h_0(t)e^{\beta_1 b_1 + \beta_2 b_2 + \cdots + \beta_m b_m}$.

In this case, we have,

$$RR = \frac{h(t, a)}{h(t, b)} = h_0(t)e^{\beta_1(a_1 - b_1) + \beta_2(a_2 - b_2) + \cdots + \beta_m(a_m - b_m)} \tag{2}$$

This value is a relative ratio indicator. It is a multiple of the risk of the occurrence of event A of the m factors $x_1, x_2, \cdots, x_m$ take the values $a_1, a_2, \cdots, a_m$ on the risk of the occurrence of event A of the m factors $x_1, x_2, \cdots, x_m$ take the values $b_1, b_2, \cdots, b_m$. Therefore, the variable RR reflects the impact of various factors on the survival time.

In many cases, the effects on the lifetime of a factor x_i are related to the state of another factor x_j. In these cases, we can say that there exists an interaction between x_i and x_j [5]. It is very important to analyze the interactions between the factors for the correct understanding of the mechanism of the influencing factors on the survival time. However, many practitioners do not know how to analyze the interaction between factors in their follow-up studies. In many textbooks on the multivariate analysis,

the methods in correctly analyzing the interaction between factors are usually not described. In this paper, we will present a method to analyze the interaction between the factors by the application of the Cox proportional hazards model. This method is applicable to any fields of multivariate data analysis when time is an observed variable.

The organization of the paper is as follows.

In the following 3 sections we describe our new method to analyze the interaction between the factors by the application of the Cox proportional hazards model. In Section 2 we give two applicable methods, product term method and dummy variable method to analyze the interaction between the factors.

In Section 3 we give an example of the applications of the methods presented in Section 2.

Some concluding remarks are in Section 4.

II.THE METHODS TO ANALYZE THE INTERACTIONS USING THE COX MODEL

Let x_1 and x_2 be the two factors affecting the survival time. Without loss of generality, we can assume the two variables are both binary variables taking values 0 or 1. By the definitions of variance and covariance [7], we know that,

If $z = ax_1 + bx_2$, then

$$Var(z) = a^2 Var(x_1) + b^2 Var(x_2) + 2abCov(x_1, x_2) \tag{3}$$

If $z = x_1 - x_2$, then

$$Var(z) = Var(x_1) + Var(x_2) - 2Cov(x_1, x_2) \tag{4}$$

where a and b are constants; $Var(x)$ is the variance of x and $Cov(x, y)$ is the covariance of x and y.

To analyze the interactions between the variables x_1 and x_2, we will discuss the following two methods of applying the Cox model.

A.Product Term Method

Let $x_3 = x_1 x_2$ be another variable. We will build a Cox model of x_1, x_2, x_3 as follows.

$$h(t, x) = h_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3} \tag{5}$$

Let $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ be the estimation of the parameters $\beta_1, \beta_2, \beta_3$ in the formula (5) by actual research data:

$$h(t, x) = h_0(t) e^{\widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3}$$

The amount of modification of x_2 to x_1 can be estimated by the following method:

Step 1: Compute the ratio of the risk of $x_1 = 1$ event A occurs to the risk of $x_1 = 0$, when $x_2 = 0$. From formula (2) and (5), we have:

$$RR_0 = e^{\widehat{\beta}_1} \tag{6}$$

Step 2: Compute the ratio of the risk of $x_1 = 1$ event A occurs to the risk of $x_1 = 0$, when $x_2 = 1$. From formula (2) and (5), we have:

$$RR_1 = e^{\widehat{\beta}_1 + \widehat{\beta}_3} \tag{7}$$

Then, the amount of changes of x_1 on the survival time due to the different state of x_2 can be expressed as

$$|RR_1 - RR_0| = |e^{\widehat{\beta}_1 + \widehat{\beta}_3} - e^{\widehat{\beta}_1}|$$

It is also called the effect of x_1 on the survival time modified by x_2.

The parameters $\beta_1, \beta_2, \beta_3$ in the formula (5) are normally estimated based on the research data. Therefore, we have to calculate its 95% confidence interval. If 0 is not contained in this interval, then we can conclude that there is an interaction between x_1 and x_2 [3]. The 95% confidence interval of $|RR_1 - RR_0|$ can be expressed as:

$$(e^{\widehat{\beta}_1 + \widehat{\beta}_3} - e^{\widehat{\beta}_1}) \pm 1.96 \sqrt{Var(e^{\beta_1 + \beta_3} - e^{\beta_1})} \tag{8}$$

In above formula, $\sqrt{Var(e^{\beta_1 + \beta_3} - e^{\beta_1})}$ can be estimated as follows:

The Taylor expansion of $\sqrt{Var(e^{\beta_1 + \beta_3} - e^{\beta_1})}$ can be written as [4]:

$$\begin{aligned} e^{\beta_1 + \beta_3} - e^{\beta_1} &= (e^{\widehat{\beta}_1 + \widehat{\beta}_3} - e^{\widehat{\beta}_1})\beta_1 + (e^{\widehat{\beta}_1 + \widehat{\beta}_3})\beta_3 \\ &+ (e^{\widehat{\beta}_1 + \widehat{\beta}_3} - e^{\widehat{\beta}_1})(1 - \widehat{\beta}_1) - (e^{\widehat{\beta}_1 + \widehat{\beta}_3})\widehat{\beta}_3 \end{aligned}$$

If we set $c = (e^{\widehat{\beta}_1 + \widehat{\beta}_3} - e^{\widehat{\beta}_1})(1 - \widehat{\beta}_1) - (e^{\widehat{\beta}_1 + \widehat{\beta}_3})\widehat{\beta}_3$, then c is a constant.

From the formula (3), we then have,

$$\begin{aligned} &Var((e^{\widehat{\beta}_1+\widehat{\beta}_3}-e^{\widehat{\beta}_1})\beta_1+(e^{\widehat{\beta}_1+\widehat{\beta}_3})\beta_3) \\ &=(e^{\widehat{\beta}_1+\widehat{\beta}_3}-e^{\widehat{\beta}_1})^2Var(\beta_1)+(e^{\widehat{\beta}_1+\widehat{\beta}_3})^2Var(\beta_3) \\ &+2(e^{\widehat{\beta}_1+\widehat{\beta}_3}-e^{\widehat{\beta}_1})e^{\widehat{\beta}_1+\widehat{\beta}_3}Cov(\beta_1,\beta_3) \end{aligned} \tag{9}$$

In above formula (9), the values of $Var(\beta_1)$, $Var(\beta_3)$ and $Cov(\beta_1,\beta_3)$ can be obtained from the variance and covariance matrices of the parameter estimations.

B.Dummy Variable Method

There are 4 different combinations of the variables x_1 and x_2 when they take values of 0 and 1:

$$(0,0);(0,1);(1,0);(1,1)$$

We then set 3 dummy variables z_1, z_2 and z_3 to represent these 4 combinations:

$$z_1=\begin{cases}1 & (x_1,x_2)=(1,1)\\ 0 & otherwise\end{cases}$$

$$z_2=\begin{cases}1 & (x_1,x_2)=(0,1)\\ 0 & otherwise\end{cases}$$

$$z_3=\begin{cases}1 & (x_1,x_2)=(1,0)\\ 0 & otherwise\end{cases}$$

Now, the Cox model becomes,

$$h(t,z)=h_0(t)e^{\beta_1z_1+\beta_2z_2+\beta_3z_3} \tag{10}$$

We can estimate the parameters β_1,β_2,β_3 in the formula (8) based on the our research data as $\widehat{\beta}_1,\widehat{\beta}_2,\widehat{\beta}_3$. Then we have,

$$h(t,z)=h_0(t)e^{\widehat{\beta}_1z_1+\widehat{\beta}_2z_2+\widehat{\beta}_3z_3}$$

The amount of modification of x_2 to x_1 can be estimated by the following method:

Step 1: Since the two cases of $(x_1,x_2)=(1,0)$ and $(x_1,x_2)=(1,1)$ are corresponding to the two cases of $(z_1,z_2,z_3)=(0,0,1)$ and $(z_1,z_2,z_3)=(1,0,0)$ respectively, the ratio of the risk of event A occurs when $(z_1,z_2,z_3)=(1,0,0)$ to the risk of event A occurs when $(z_1,z_2,z_3)=(0,0,1)$ can be computed by formula (2) and (8) as $RR_1=e^{\widehat{\beta}_1-\widehat{\beta}_3}$.

By formula (4), the 95% confidence interval of RR_1 can be expressed as:

$$e^{(\widehat{\beta}_1-\widehat{\beta}_3)\pm1.96\sqrt{Var(\beta_1)+Var(\beta_3)-2Cov(\beta_1,\beta_3)}} \tag{11}$$

Step 2: Since the two cases of $(x_1,x_2)=(0,0)$ and $(x_1,x_2)=(0,1)$ are corresponding to the two cases of $(z_1,z_2,z_3)=(0,0,0)$ and $(z_1,z_2,z_3)=(0,1,0)$ respectively, the ratio of the risk of event A occurs when $(z_1,z_2,z_3)=(0,1,0)$ to the risk of event A occurs when $(z_1,z_2,z_3)=(0,0,0)$ can be computed by formula (2) and (8) as $RR_0=e^{\widehat{\beta}_2}$.

By formula (4), the 95% confidence interval of RR_0 can be expressed as:

$$e^{\widehat{\beta}_2\pm1.96\sqrt{Var(\beta_2)}} \tag{12}$$

We can conclude that there must be an interaction between x_1 and x_2 if the 95% confidence interval of RR_0 and the 95% confidence interval of RR_1 has no intersection. Therefore, the amount of modification of x_2 to x_1 can be evaluated by the difference of RR_1 and RR_2.

III.APPLICATIONS OF THE METHODS

In order to investigate whether new treatments can improve survival in patients with malignant, we have recorded their treatment method x_1 ($x_1=0$ corresponding to the new treatment method of treatment; $x_1=0$ corresponding to a traditional method of treatment) and their lymph node metastasis x_2 ($x_2=1$ corresponding to lymph node metastasis; $x_2=0$ corresponding to no lymph node metastasis), for pathological diagnosis of 63 patients with malignant. We have also made a follow-up observation of their survival time ($time$). If we represent the random event "death" as a random variable $outcome$, then $outcome=0$ for the patients have not died at the end of the follow-up period, otherwise $outcome=1$.

TABLE 1 The Survival Time and Its Influencing Factors of 63 Cases of Malignant Patients

No.	x_1	x_2	$time$	out	No.	x_1	x_2	$time$	out
1	1	0	52	1	33	0	0	120	1
2	0	1	51	1	34	1	1	40	1
3	1	1	35	1	35	0	0	26	0
4	1	0	103	1	36	0	1	120	1
5	0	1	7	0	37	1	1	120	1
6	0	1	60	1	38	0	0	120	0
7	0	1	58	1	39	0	0	3	0
8	1	1	29	1	40	0	0	120	0
9	1	1	70	1	41	0	1	7	0
10	0	1	67	1	42	0	1	18	1
11	0	1	66	1	43	1	1	120	1
12	1	0	87	1	44	0	1	120	1
13	1	0	85	1	45	0	1	15	1
14	0	1	82	1	46	0	0	4	0
15	1	1	76	1	47	1	1	120	1
16	1	1	74	1	48	0	0	16	0
17	1	1	63	1	49	0	1	24	0
18	0	0	101	1	50	0	0	19	0
19	0	0	100	1	51	0	1	120	1
20	1	1	66	1	52	0	0	24	0
21	0	0	93	1	53	0	0	2	0
22	1	1	24	1	54	0	1	120	1
23	1	0	93	1	55	1	1	12	1
24	1	0	90	1	56	0	0	5	0
25	1	1	15	1	57	0	0	120	1
26	0	0	3	0	58	1	1	120	1
27	1	0	87	1	59	0	0	7	0
28	0	0	120	0	60	0	0	40	0
29	0	0	120	0	61	1	1	108	1
30	0	0	120	0	62	0	1	24	1
31	0	1	120	1	63	0	0	16	0
32	1	1	120	1					

TABLE 2 The Maximum Likelihood Estimation of the Cox Model Parameters for 63 Cases of Patients with Malignant

	Estimations	SE of β	P values
Treatment method x_1	1.591	0.605	0.009
Lymph node metastasis x_2	1.353	0.522	0.009
$x_3 = x_1x_2$	-1.508	1.702	0.032

TABLE 3 The Variance and Covariance Matrix of Parameters

	x_1	x_2	x_3
x_1	0.366	0.201	-0.365
x_2	0.201	0.272	-0.272
x_3	-0.365	-0.272	0.493

The follow-up results for the 63 patients are shown in Table 1.

We input the data in Table 1 into the computer, and then performed a Cox model fitting by using the process phreg of statistical package SAS 9.0 [1]. We obtained a fitting result of Table 2, and the variance and covariance matrices of parameters of Table 3.

From formula (6) we can see: in the cases of no lymph node metastasis ($x_2 = 0$), the risk of death in patients with traditional method of treatment ($x_1 = 1$) is about $RR_0 = e^{\widehat{\beta}_1} = e^{1.591} = 4.909$ times for those with new treatment method ($x_1 = 0$).

From formula (7) we can see: in the cases of lymph node metastasis ($x_2 = 1$), the risk of death in patients with traditional method of treatment ($x_1 = 1$) is about $RR_1 = e^{\widehat{\beta}_1+\widehat{\beta}_3} = e^{1.591-1.508} = 1.086$ times for those with new treatment method ($x_1 = 0$).

$$|RR_1 - RR_0| = 4.909 - 1.086 = 3.823$$

From formula (9) we know,

$$\begin{aligned}
&Var(RR_1 - RR_0)\\
&= (e^{\widehat{\beta}_1+\widehat{\beta}_3} - e^{\widehat{\beta}_1})^2 Var(\beta_1) + (e^{\widehat{\beta}_1+\widehat{\beta}_3})^2 Var(\beta_3)\\
&+2(e^{\widehat{\beta}_1+\widehat{\beta}_3} - e^{\widehat{\beta}_1})e^{\widehat{\beta}_1+\widehat{\beta}_3} Cov(\beta_1, \beta_3)\\
&= 3.823^2 \times 0.366 + 1.086^2 \times 0.493\\
&+2 \times 3.823 \times 1.086 \times (-0.365) = 2.899
\end{aligned}$$

Then we know that the 95% confidence interval of $|RR_1 - RR_0|$ must be

$$3.823 - 1.96\sqrt{2.899} \sim 3.823 + 1.96\sqrt{2.899}$$

In other words, the 95% confidence interval of $|RR_1 - RR_0|$ is $0.486 \sim 7.160$.

IV.CONCLUDING REMARKS

This paper has introduced two applicable methods, product term method and dummy variable method to analyze the interaction between the factors by the application of the Cox proportional hazards model. An example presented in Section 3 shows that the two methods are very practical.

The result of this application shows that in the case of no lymph node metastasis the new treatment method is better. While for the case of lymph node metastasis, the risk of death with new treatment method is not much lower than that of with the traditional treatment method. From the confidence interval of view, this difference reaches a statistical significance. This shows that there is an interaction between lymph node metastasis and treatment methods. The efficacy of the new treatment method is modified by the severity of the disease.

ACKNOWLEDGEMENT

This work was supported in part by the Natural Science Foundation of Fujian (Grant No.2013J01247), and Fujian Provincial Key Laboratory of Data-Intensive Computing and Fujian University Laboratory of Intelligent Computing and Information Processing.

REFERENCES

[1] Ron Cody, SAS Statistics by Example, New York, SAS Publishing, 2011.

[2] D. R. Cox, Analysis of Survival Data. New York: Chapman and Hall 1984.

[3] E. Frank, Jr. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, New York, Springer-Verlag, 2010.

[4] R. J. Baugh, Foundations of Mathematical Analysis, New York, Dover Publications, 2010.

[5] M. H. Katz, Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers,3 edition, Cambridge, Cambridge University Press, 2011.

[6] E. T. Lee, Statistical Methods for Survival Data Analysis, 4th Edition. New York: John Wiley Sons, Inc. 2013.

[7] J. T. McClave, Statistics, 12 edition, New York, Pearson, 2012.