ESTEEM, Vol. 4, No. 2, 2008, 3–14

Using Kaplan Meier and Cox Regression in Survival Analysis: An Example

Teoh Sian Hoon

ABSTRACT

The Kaplan Meier procedure is used to analyze data based on the survival time. This article provides an example on how to use Kaplan Meier and Cox regression with three objectives. The objectives are finding the percentage of survival at any time of interest, comparing the survival time of two studied groups and examining the effect of continuous covariates with the relationship between an event and possible explanatory variables. The example is discussed based on the breast cancer survival dataset from Statistical Package for the Social Sciences (SPSS).

Keywords: *Survival time, the Kaplan Meier procedure, Cox regression*

Introduction

This article provides an example on how to use the Kaplan Meier procedure and Cox regression to analyze data with censored data on survival time as well as to find the relationship between an event and possible explanatory variables. The procedures in the Kaplan Meier and the Cox regression were reviewed in the following literature review. Using the Cox regression model to find the effects of covariates requires the use of a statistical software package because straightforward single equation for the estimation is not available (Daniel, 2005). In this paper, the software package SPSS (Statistical Package for the Social Sciences) was used. After the data are included in the analysis using SPSS, the data are analyzed based on the procedures.

Literature Review

Survival analysis is used to analyze data corresponding to survival time. Survival time is the time taken when an end event occurs in the data set.

Thus, it is the time to events, and is also known as failure-time data or the end point. Time to events include survival time in medical events and non-medical events. An example of survival time in a medical event is the survival time until death or until being discharged from hospital. Another good example is the time for a tooth experiences carries as discussed in a longitudinal oral health study (Komárek, Lesaffre, Härkänen, Declerck & Virtanen, 2005; Lesaffre, 2005). On the other hand, an example of survival time in non-medical event is the time from graduation until one gets a job. A famous example of the application of survival analysis for non-medical event is finding the determinants of the survival of a network in franchising, where time is an important variable for the development of franchising (Perrigot, Cliquet and Mesbah, 2004). All the cases in the above examples have data consist of censored observations in which the end event has not happened in every observation or when information on a case is only known for a limited duration. The censoring time is the main information to find cumulative survival probability in the survival analysis. In other words, the great advantage of using survival analysis is to analyze censored cases in analysis.

The Kaplan-Meier procedure (Kaplan & Meier, 1958) is used to calculate the survival rate from the survival function. It involves estimating the probability of surviving for a specified length of time. The advantage of using Kaplan-Meier curves is that they are non-parametric, where no assumptions are made on the distribution of survival times (Kaplan & Meier, 1958; Daniel, 2005). The survival function is the number of individuals with survival time, which is at least t time periods divided by the number of individuals in the study. The Kaplan-Meier estimate of the survival function is a product-limit estimate as indicated in Equation 1,

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \tag{1}$$

where n_j = number of individuals alive just before time $t(_j)$, and d_j = number of deaths at $t(_j)$ for $t(_k) \leq t \leq t(_{k+1})$, and $k = 1, 2, \ldots, r$.

The two useful regression models, namely linear regression models and logistic regression model are used for continuous outcome measures and binary outcome measures. Cox regression or proportional hazard regression is an additional type of regression models. It is used when the dependent observations consist of a mixture of either time-until-event data or censored time observations (Daniel, 2005). The function involved is hazard function, which describes the conditional probability of an

individual who has survived to time *t* and will die in the next small period of time. The Cox proportional hazard model assumes that the hazards for two groups are proportional (Collet, 2003). The regression model is described in Equation 2.

$$h(t, x_i) = e^{f(x_i)} h_0(t) \tag{2}$$

where $e^{f(x_i)}$ indicates how different covariates affect survival (i.e., compares the hazard to the baseline) and $h_o(t)$ is an arbitrary baseline hazard function that is assumed to be the same for all groups.

By rearranging the Equation 2, we can get the exponentiated coefficient, which represents the hazard ratio for the basis in proportional hazard regression as indicated in Equation 3.

$$\frac{h(t, x_i)}{h_0(t)} = e^{f(x_i)} \tag{3}$$

Objectives

There are three main objectives in the following discussion. They are (1) to find the percentage of survival at any time of interest, (2) to compare the survival time of two studied groups, and (3) to examine the effects of continuous covariates. The Kaplan Meier procedure was used for the first and second objectives. Cox regression was used for the third objective.

Methodology

SPSS was used in this analysis. Kaplan Meier and Cox regression are the two main analyses in this paper. The Kaplan Meier procedure is used to analyze on censored and uncensored data for the survival time. It is also used to compare two treatment groups on their survival times. The Kaplan Meier technique is the univariate version of survival analysis. To present more details in the survival analysis, further analysis using Cox regression as multivariate analysis is presented. Cox regression allows the researcher to include predictor variables (covariates) into the models. Cox regression will handle the censored cases correctly. It will provide estimated coefficients for each of the covariates that allow us to assess the impact of multiple covariates in the same model. We can also use Cox regression to examine the effect of continuous covariates. The

steps required in SPSS to perform the above objectives are listed as follows.

Variables Used

The event of interest in a medical research using survival analysis is death due to a disease. There are two groups of status, which are censored data and uncensored data. The occurrence of censored observations may due to a few reasons. Firstly, the observations are still alive at the end of study for which the critical event has not yet occurred. Secondly, the observations' follow-up information are lost. This can be caused by the person's reluctance to turn up for the following study days after committing to the study. Thirdly, the event occurs but the cause is unrelated to the disease.

The status of data is recorded to identify whether the observation is a censored data or an uncensored data. For censored data, the status is denoted as '0'. For uncensored data, the case from an event of dying from the disease is denoted as '1'. Normally a factor is used to indicate whether the observation is in the treatment group or control group. If there are only one treatment group and one control group, the factor is set as '0' for control group and '1' for the treatment group. But, if there are two treatment groups and one control group, '0' is set for the control group, '1' is set for the first treatment, and '2' is set for the second treatment.

The event of interest in this study is death due to breast cancer. A few variables were used in this discussion. Firstly, status of a variable was used to indicate the status of censored data or uncensored data. Status of '0' denotes censored data and status '1' denotes uncensored data. Secondly, variable time was used to indicate time of occurrence for censored and uncensored observations. Thirdly, variable Lymph Nodes was included as a factor to compare the survival times of two groups. In this example, factor Lymph Nodes has two categories. They are status of 'No' for not having Lymph Nodes and status of 'Yes' for having Lymph Nodes. In the data set, '0' was recorded to define the status of 'No' and '1' was recorded to define the status of 'Yes' for the factor Lymph Nodes. Fourthly, variables age, Histologic Grade and Lymph Nodes were used as covariates in a further analysis, namely Cox regression.

Steps for the First and Second Objectives: Using the Kaplan Meier Procedure

Kaplan Meier is used to analyze survival time data. The following steps will give descriptive statistics for the survival time and a survival plot for the survival function. Step 1 to step 4 are checked on to perform the first objective. Additional two steps, namely step 5 and step 6, are required to perform the second objective.

Step 1: From Menu bar, click on 'Analyze', point to 'Survival', followed by 'Kaplan Meier'.

Step 2: Select and move variable 'time' to the 'Time' box and variable 'status' to the 'Status' box.

Step 3: Click 'Define Event' under 'Status' box, then, include '1' as an event as defined and click 'Continue'.

Step 4: Click 'Option' button and mark on 'Mean and median survival' under 'Statistics' dialogue box; mark on 'Survival' under 'Plots' dialogue box.

Step 5: Select and move variable 'Lymph Node' to the 'Factor' option.

Step 6: Click 'Compare Factor...' button and click on 'Log rank'. Then click 'Continue' button.

Steps for the *Third* Objective: Using Cox Regression

We can determine whether the two groups differ with a few predictor variables, namely Lymph Node, Histologic Grade and age by performing Cox regression. The independent variables (covariates) can be continuous or categorical; for categorical variables, reference groups should be indicated. By default the last group is referred as the reference group. In this example, Histologic Grade and Lymph Nodes are categorical variables, the reference category for Histologic Grade is "3", and the reference category for Lymph Nodes is "1". The following steps will give the estimated variables in Cox regression.

Step 1: From Menu bar, click on 'Analyze', point to 'Survival', followed by 'Cox Regression'.

Step 2: Select and move variable 'time' to the 'Time' box and variable 'status' to the 'Status' box.

Step 3: Click 'Define Event' under 'Status' box, then, include '1' as an event as defined and click 'Continue'.

Step 4: Select and move variable 'Lymph Node', 'Histologic Grade' and 'age' to the 'Covariates' box.

Step 5: Click 'Categorical...' button and include variables 'Lymph Node' and 'Histologic Grade' into 'Categorical Covariates' box. Then, click 'Continue'.

Step 6: Click 'Options...' button and click on 'CI for exp(B)' checkbox in 'Cox Regression: Options' dialogue box.

Results

The following results are presented according to the three objectives.

Using the Kaplan Meier Procedure

Table 1, Table 2, and Figure 1 are presented as below for the analysis based on the first objective. Table 1 shows the number of events, namely the number of cases is 72, with the percentage of censored cases being 94%. Table 2 shows the mean of survival time is 122.692 months, with the standard error of 1.307 months. Figure 1 shows the survival plots. It is shown in the diagram that at 70 months, 90% of the observations were still alive. From Figure 1, more information on the percentage of the survival for different months can be accessed by referring to the specific month and looking for the associate survival rate.

Table 1: Case Processing Summary

		Censored	
Total *N*	*N* of events	*N*	Percent
1,207	72	1,135	94.0%

Table 2: Means for Survival Time

Mean[a]			
		95% Confidence interval	
Estimate	Std. error	Lower bound	Upper bound
122.692	1.307	120.131	125.253

[a] Estimation is limited to the largest survival time if it is censored.

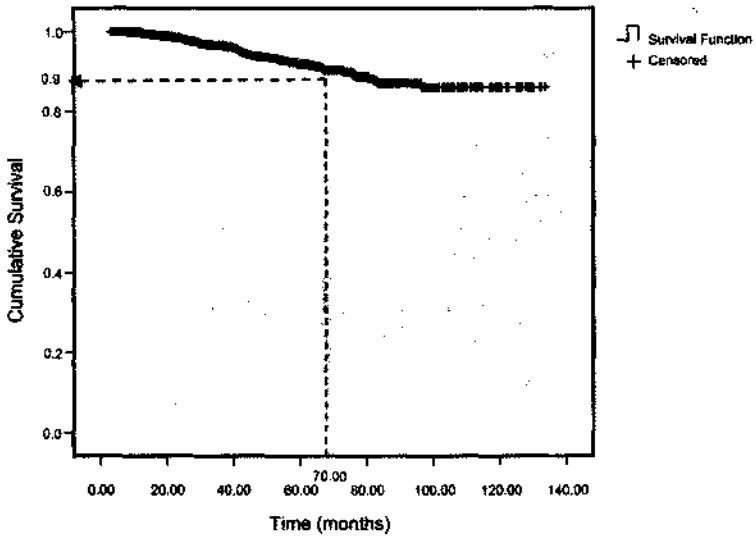

Figure 1: Survival Function

Table 3, Table 4, Table 5 and Figure 2 are presented as below for the analysis on the second objective. Table 3 shows the number of cases for the two categories in Lymph Nodes, with cases of 'No' at 929 observations and cases of 'Yes' at 278 observations. Thus, there are 929 observations which do not have Lymph Nodes and 278 observations which have Lymph Nodes. Table 4 shows the mean survival times for the two groups, with the mean for cases of 'No' (without Lymph Nodes) as 124.920 months and the mean for cases of 'Yes' (with Lymph Nodes) as 111.331 months. Table 5 shows the results of log-rank test with the p-value of .000, which indicates that there is a significant difference between the two groups on having a shorter time to event. The survival plot (Figure 2) shows the group without the Lymph Nodes has a longer survival time to event compared to the group with Lymph Nodes. This scenario is shown in Figure 2, whereby 92% of patients without Lymph Nodes were still alive at 60 months as compared to 82% of patients with Lymph Nodes. From Figure 2, more information on the survival rate for different months for the two groups can be retrieved by referring to the specific month and looking for the associated survival rates.

Table 3: Case Processing Summary

yes or no (with Lymph Nodes)	Total *N*	*N* of events	Censored	
			N	Percent
No	929	42	887	95.5%
Yes	278	30	248	89.2%
Overall	1,207	72	1,135	94.0%

Table 4: Means for Survival Time

yes or no (with Lymph Nodes)	Mean[a]			
			95% Confidence interval	
	Estimate	Std. error	Lower bound	Upper bound
No	124.920	1.400	122.177	127.664
Yes	111.331	3.008	105.436	117.226
Overall	122.692	1.307	120.131	125.253

Table 5: Overall Comparisons

	χ^2	*df*	Sig.
Log Rank (Mantel-Cox)	15.988	1	.000

Note: Test of equality of survival distributions for the different levels of Lymph Nodes.

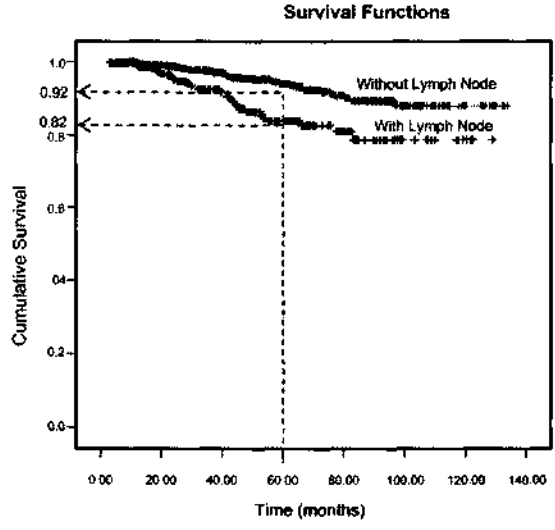

Figure 2: Survival Plot for Comparison of the Two Groups

Results from the Cox regression are presented in Table 6, Table 7, Table 8, and Table 9. Table 6 shows that only 76.2% of the observations or cases are available in the analysis and the number of cases dropped is at 23.8%, namely there are 287 cases reported as missing data.

From Table 7, we notice that the reference category for histgrad (Histologic Grade) is '3'. The category is indicated by value '0' both in the codes (1) and (2). On the other hand, the category for histgrad of '2' is indicated by values '0' and '1' in the codes (1) and (2) respectively. Then, the category for histgrad of '1' is indicated by values '1' and '0' in the codes (1) and (2) respectively. For Lymph Nodes, the reference category is '1' with values '0' in the code (1). Category '0' in the Lymph Nodes has value'1' in the code (1).

Table 6: Case Processing Summary

Case	*N*	Percent
Cases available in analysis		
Event[a]	56	4.6%
Censored	864	71.6%
Total	920	76.2%
Cases dropped		
Cases with missing values	287	23.8%
Cases with negative time	0	.0%
Censored cases before the earliest event in a stratum	0	.0%
Total	287	23.8%
Total	1,207	100.0%

[a] Dependent variable: Time (months).

Table 7: Categorical Variable Codings[c,d]

Variable	Frequency	(1)[a]	(2)
histgrad[b]			
1 = 1	79	1	0
2 = 2	514	0	1
3 = 3	327	0	0
yes_or_no[b]			
0 = No	692	1	
1 = Yes	228	0	

[a] The (0,1) variable has been recoded, so its coefficients will not be the same as for indicator (0,1) coding. [b] Indicator Parameter Coding. [c] Category variable: histgrad (Histologic Grade).
[d] Category variable: yes_or_no (Lymph Nodes).

Table 8 shows the model is significant with chi square, χ^2 value of 18.191 and *p*-value less than .05. Table 9 provides the *p*-values and the hazard ratio (Exp(B)) of the variables. All *SE* values in Table 9 are small, and the problem of multicolinearity is under controlled. For the confounder model, the most important variable to be looked into is the group factor, which is the Lymph Nodes. The result shows the *p*-value is .012, which is significant as reported in the Kaplan Meier analysis. The associate hazard ratio (HR) as indicated in Exp(B) is .5, which is less than '1'. For reporting HR, there are three possibilities: (a) a value of '1' means there is no differences between two groups in having a shorter time to event, (b) a value of 'more than 1' means that the group of interest is likely to have a shorter time to event as compared to the reference group, and (c) a value of 'less than 1' means that the group of interest less likely to have a shorter time to event comparing to the reference group. Therefore, the group of interest for Lymph Nodes (which is '0' – without lymph node) is less likely to have a shorter time to event (death) as compared to the reference group. Table 9 also shows that only 'Lymph Nodes' has significant result, whereas other variables have insignificant results.

Table 8: Omnibus Tests of Model Coefficients [a,b]

-2 Log likelihood	Overall (score)			Change from previous step			Change from previous block		
	χ^2	*df*	Sig.	χ^2	*df*	Sig.	χ^2	*df*	Sig.
663.850	18.191	4	.001	16.943	4	.002	16.943	4	.002

[a] Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 680.793. Beginning Block Number 1. Method = Enter.

Table 9: Variables in the Equation

Variable	B	*SE*	Wald	*df*	Sig.	Exp(B)	95.0% CI for Exp(B) Lower	Upper
Age	−.013	.011	1.601	1	.206	.987	.966	1.007
Histgrad			4.978	2	.083			
Histgrad(1)	−1.147	.737	2.426	1	.119	.317	.075	1.345
Histgrad(2)	−.521	.278	3.514	1	.061	.594	.344	1.024
ln_yesno	−.692	.277	6.266	1	.012	.500	.291	.861

Interaction of variables can be performed if more than one variable has a significant result after detecting that the main effect is significant. Interactions will offer more details about particular results for the categories. In this example, only the group factor (Lymph Nodes) shows significant difference on the survival rate. Thus, analysis for interaction of the variables is not included.

Conclusion

The Kaplan Meier procedure is used to analyze survival time data for censored and uncensored observations. In addition, it is used to compare two treatment groups on the survival time. The Kaplan Meier is a univariate analysis and further analysis for multivariate analysis can be done by using Cox regression. Cox regression presents a more realistic situation. Therefore, predictors for a shorter survival time to death can be detected.

Acknowledgment

SPSS dataset, screenshots and applications are used by permission of SPSS.

References

Collet, D. (2003). *Modelling survival data in medical research.* Boca Raton, FL: Chapman & Hall/CRC.

Daniel, W. W. (2005). *Biostatistics: A foundation for analysis in the health sciences.* River Street, U.S.: John Wiley & Sons, Inc.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53,* 457–481.

Komárek, A., Lesaffre, E., Härkänen, T., Declerck, D., & Virtanen, J. I. (2005). A Bayesian analysis of multivariate doubly-interval-censored dental data. *Biostatistics, 6*(1), 145–155.

Lesaffre, M. (2005). An overview of methods for interval-censored data with an emphasis on application in dentistry. *Statistical Methods in Medical Research, 14*(6), 539–552.

Perrigot, R., Cliquet, G., & Mesbah, M. (2004). Possible applications of survival analysis in franchising research. *The International Review of Retail, Distribution and Consumer Research, 14*(1), 129–143.

TEOH SIAN HOON, Department of Information Technology and Quantitative Sciences, Universiti Teknologi MARA Pulau Pinang, 13500 Permatang Pauh, Pulau Pinang, MALAYSIA. E-mail: teohsian@ppinang.uitm.edu.my