RESEARCH ARTICLE | AUGUST 27 2024

Survival analysis using censored lung cancer data: A preliminary study

Siti Afiqah Muhamad Jamil ✉; Nurul Izzah Ali; Mahayaudin M. Mansor; Ahmad Zia Ul-Saufi; Muhammad Ammar Shafi

AIP Conf. Proc. 3123, 030006 (2024)
https://doi.org/10.1063/5.0224907

View Online

Export Citation

Survival Analysis using Censored Lung Cancer Data: A Preliminary Study

Siti Afiqah Muhamad Jamil[1, a)], Nurul Izzah Ali[1, b)], Mahayaudin M Mansor[1, c)], Ahmad Zia Ul-Saufi[1, d)] and Muhammad Ammar Shafi[2, e)]

[1]School of Mathematical Sciences, College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia
[2]Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, 84600,
Pagoh, Johor Malaysia

[a)] Corresponding author: afiqahjamil@uitm.edu.my
[b)] izzah067@gmail.com
[c)] maha@uitm.edu.my
[d)] ahmadzia101@uitm.edu.my
[e)] ammar@uthm.edu.my

Abstract. The primary goal of Exploratory Data Analysis (EDA) is exploring and understanding data in order to gain insights and guide for further analysis. It allows for data cleaning which involves removing redundancies, handling missing values, correcting errors and transforming data if necessary as it is an important part of overall data preparation process. The aim of this study is to conduct a preliminary study before applying the survival method of analysis to censored lung cancer observations. The Kaplan-Meier survival curve, proportional hazard assumption, time varying covariate assumption via Scaled Schoenfeld residuals, Cox-Snell residuals for overall goodness of fit of the model assumption, and normality assumption via quantile-quantile (Q-Q) plot were all used in this study. The study discovered that the lung cancer censored observations were not violated with the parametric assumption among semi-parametric, and non-parametric assumptions. Thus, future work is recommended to include a comparison of the parametric method of survivals using the lung cancer data. The application of survival analysis would be ambiguous and mislead the researcher if the fundamental part of survival being left out. As a result, this study may aid in identifying appropriate assumptions for prior application on the survival analysis of censored observations.

INTRODUCTION

Exploratory Data Analysis (EDA) is the preliminary investigation of data using descriptive statistics and graphical representations to uncover patterns, detect anomalies, test hypotheses, and verify assumptions. Because survival analysis is a well-known statistical method for analysing clinical data that is also frequently used in medical research, EDA could help to simplify the understanding of a dataset's structure and make data modelling relatively easy [1]. It also assists us in understanding the interaction between variables, providing us with a broader perspective on the information and allowing us to develop on it by effectively utilising the interactions between variables. It also makes it easier to measure statistical approaches in the dataset.

[2] study compared parametric and semiparametric survival analysis among HIV-infected adults but did not focus on the assumptions related to survival analysis. Aside from mentioning the assumption prior to the survival analysis, similar studies were found on the application of survival analysis related to exploratory data with descriptive analysis and overall survival on time to event data [3, 4, 5].

Despite the use of previous survival analysis studies, the flow of the preliminary study involving survival analysis was not fully explained. Aside from that, little progress has been made on the statistical write-up for survival analysis.

The 6th ISM International Statistical Conference (ISM-VI) 2023
AIP Conf. Proc. 3123, 030006-1–030006-10; https://doi.org/10.1063/5.0224907
Published under an exclusive license by AIP Publishing. 978-0-7354-5030-1/$30.00

As a result, the goal of this study is to conduct a preliminary study before applying the survival method of analysis to censored lung cancer observations.

METHODOLOGY

The preliminary study included non-parametric, semi-parametric, and parametric regression models. These methods would ensure that the analysis chosen for the data is precise and that the assumptions related to the survival analysis have been correctly checked, resulting in a good, precise output and reducing the biases.

Procedure in Exploratory Data Analysis

This study only focused on exploratory data analysis to ensure that the appropriate methods were used to analyse the data. Basically, there are three popular approaches to survival analysis: parametric, non-parametric, and semi-parametric. Testing all assumptions related to these approaches is important to ensure that the best method chosen is appropriate for the lung cancer data.

To begin, this study had to examine the Kaplan-Meier (K-M) survival plot and the log-rank test for categorical variables for the non-parametric model. The K-M survival curve was created to compare the survival pattern of lung cancer patients with categorical variables such as gender, race, type of lung cancer, and treatment of lung cancer. The log-rank test is then used to address the survival curve of K-M in numerical procedures in order to observe differences in survival among groups in categorical variables.

Continuously, the assumptions for semi-parametric regression models are related to the Cox proportional hazard model, which this study had to test on the proportional hazard assumption and the time varying covariate assumption by using either Scaled Schoenfeld residuals or Unscaled Schoenfeld residuals and either Nelson-Aalen plot or log minus log respectively. Furthermore, the Cox-Snell residuals were examined for the overall goodness of fit of the model in this study.

The two plots that can be used to check on the normality of a parametric model are the quantile-quantile (Q-Q) plot and the probability (P-P) plot. Furthermore, the best fit model can be evaluated using the Cox-Sell residual.

RESULTS

Descriptive Statistics of Lung Cancer

This study included 54 patients from a general hospital in Johor Bahru who had been diagnosed with lung cancer. The descriptive summary is shown in the table below.

TABLE 1. Descriptive statistics for continuous data.

Characteristics (Continuous data)	No. (%)
Age (year)	
Mean, Standard deviation (SD)	59.85, 12.30
Median, range	61.50, (13,85)
≤ 60	26 (41.15)
> 60	28 (51.85)

According to Table 1, people over the age of 60 have a higher risk of developing lung cancer than those under the age of 60. According to the statistics, patients over the age of 60 accounted for 28 patients, while patients under the age of 60 accounted for 26 patients.

As a result, statistics revealed that the average age was 59.85, with a standard deviation of 12.30. With a median of 61.50, 50 percent of those diagnosed with lung cancer were found to be between the ages of 61.50 and 61.50. The average age of patients diagnosed with lung cancer is 13 years old, with a maximum age of 85 years old. The frequencies of each variable for categorical variables are shown in the table below.

TABLE 2. Descriptive statistics for categorical data.

Characteristics (Categorical data)	No. (%)
Types of lung cancer	
Adenocarcinoma, NSCLC	19 (35.19)
Large Cell Carcinoma, NSCLC	17 (31.48)
Small Cell lung cancer, SCLC	12 (22.22)
Squamous cell carcinoma, NSCLC	6 (11.11)
Treatment of lung cancer	
Chemoradiotherapy, CCRT	32 (59.26)
Chemotherapy, Chemo	5 (9.26)
Chemotherapy and Surgery, Chemosurgery	11 (20.37)
Chemotherapy and Targeted Therapy, ChemoTarget	5 (9.26)

Timeline of the Duration of Survival

Figure 1 depicts the time it took for lung cancer patients to survive for nine years, from February 15th, 2008 to February 15th, 2017. The time when patients finished their follow-up treatments was the event of interest in this study. The completion data, which represent the fixed duration of survival, and the right censored observations were used to calculate the duration of survival. The green bullets on the timeline with the full coloured line indicate that the patients completed their follow-up treatments and died before the study period ended. It is known as an exact observation, whereas the grey colour represents censored observations.

The timeline depicts 54 patients with 25 observations, 9 censored observations, 7 left censored observations, and 13 interval censored observations. In summary, 25 of the patients have already died, which contributes to the exact observation as the event of interest occurred, whereas the remaining 29 observations represent the censored observation.

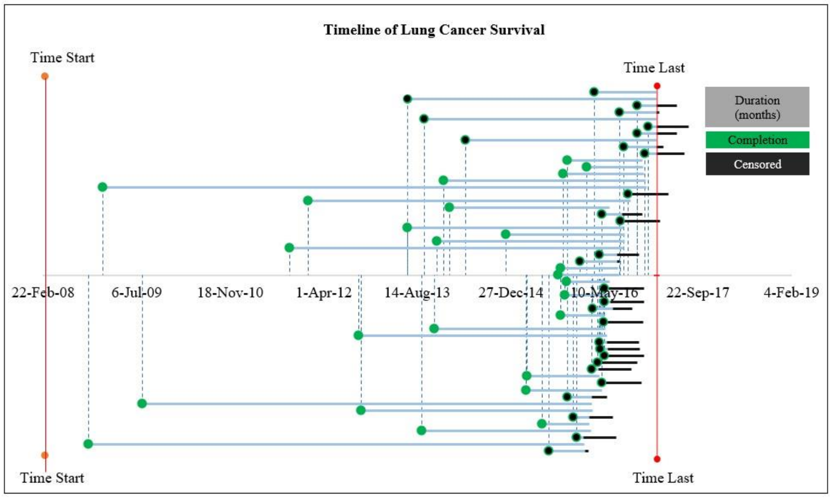
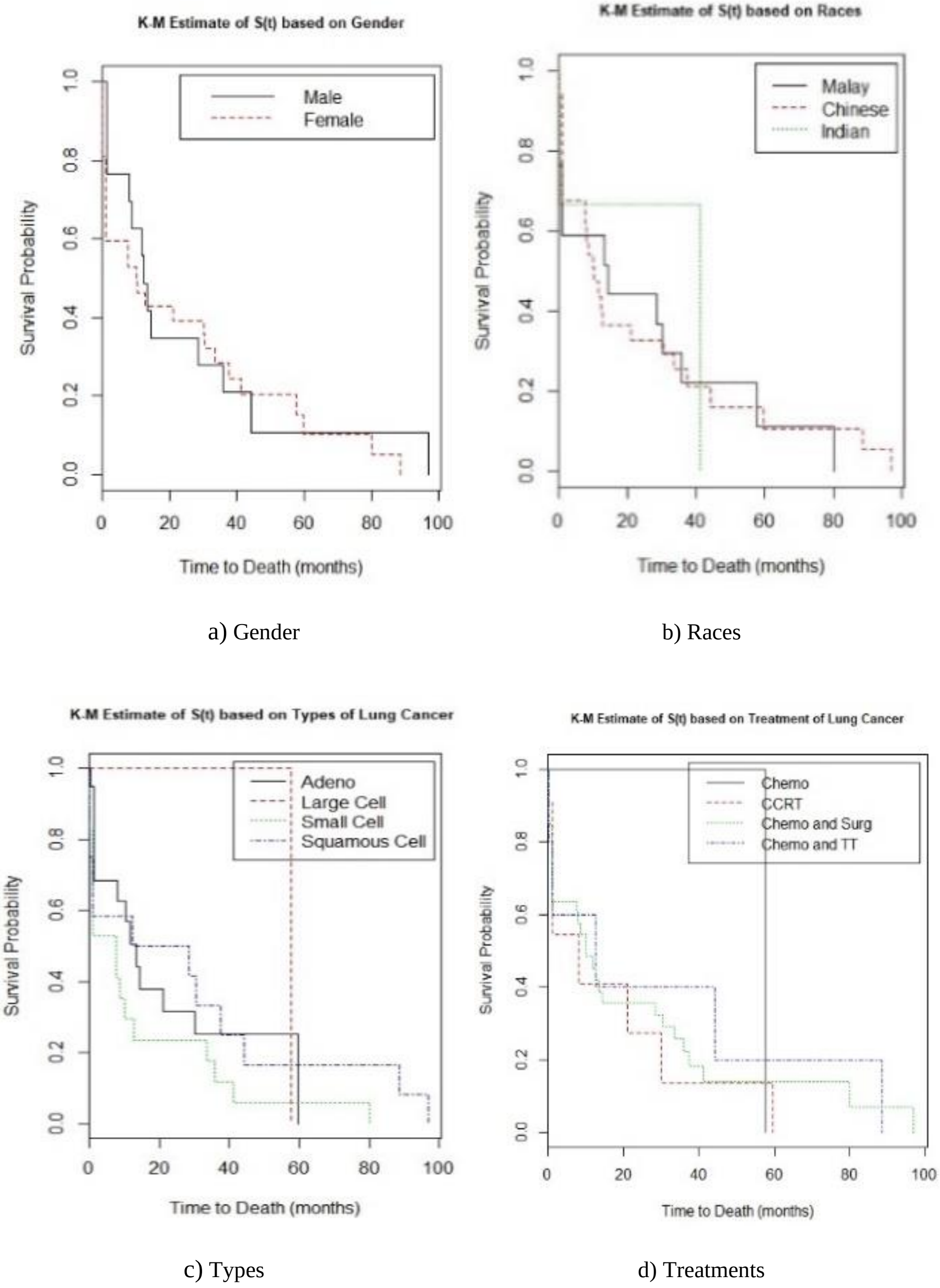


FIGURE 1. The timeline of duration of survival of lung cancer patients

Kaplan-Meier Product Limit Survival Curve and Log-Rank Test

This study must examine the survival times for categorical variables using graphical methods of the Kaplan-Meier (K-M) estimation. In addition, to test the proportional hazard assumption, this study will look for parallelism in plotting survival curves.

Figure 2 depicts a comparison of male and female genders, as well as Malay, Chinese, and Indian races, using the Kaplan-Meier (K-M) survival curve. The two curves between male and female did not show a parallel line because they were intersecting with each other, and the lines of graph are also intercepting and not parallel. Furthermore, using the K-M survival curve to compare categorical variables is difficult because the curves tend to intersect within each other and the lines between the groups of each variable are not parallel. Thus, the log-rank test results correspond to the Kaplan-Meier curves, as shown in Table 3.



a) Gender

b) Races

c) Types

d) Treatments

FIGURE 2. Comparison between the gender, races, types and treatments of lung cancer using K-Meier estimate

TABLE 3. Log-rank test's output.

Log-rank (Mantel-Cox)	*z*-value	df	*P*-value
Gender	0.919	1	0.358
Races	-0.289	2	0.773
Types	0.937	3	0.349
Treatments	0.923	3	0.356

The results of the log-rank test can be summarised as follows:

1. The male and female have equal chances of survival.
2. Survival is the same for Malay, Chinese, and Indian people.
3. Survival rates for adenocarcinoma, large cell carcinoma, small cell lung cancer, and squamous cell carcinoma are all the same.
4. Survival does not differ between chemotherapy, chemoradiotherapy, chemotherapy and surgery, or chemotherapy and targeted therapy.

Thus, Kaplan-Meier survival analysis was ineffective when applied to this lung cancer data because the assumption of parallelism among categorical variables was violated, as supported by the log-rank test.

Proportional Hazard Assumptions

Furthermore, this study must investigate the time-varying effect of the variables in order to illustrate the scenario of handling medical records data. When the proportional hazard assumption is violated, a time-varying effect emerges. As a result, after fitting the Cox proportional hazard model, this study must examine the proportional hazard assumption to identify the varying effect. The results of exploring the assumption of proportional hazard or testing for proportionality using two different models are shown in Table 4. In this study, the only continuous covariate whose form must be assessed is age. Figure 3 depicts the martingale residual plot for the variable of age in order to determine the best functional form to use for the continuous variable. The Martingale residual plot in Fig. 3 above is used to investigate the functional form that must be used to deal with the continuous variable in the lung cancer data. The blue line represents the loess pointwise confidence band, and the curvy behaviour of the loess fit indicates that this study should take into account higher order of Age. As a result, this study considers the age and square of age in the model, as shown in Table 5, because Fig. 3 yields results that require evaluating the higher order of the Age. Furthermore, the unscaled Schoenfeld residuals are based on the global results in the table above. An associated global significant test for both models yields P-values of 0.066 and 0.117, respectively, indicating that the lung cancer data is proportionally fit to the model. To put it another way, both models passed the global test.

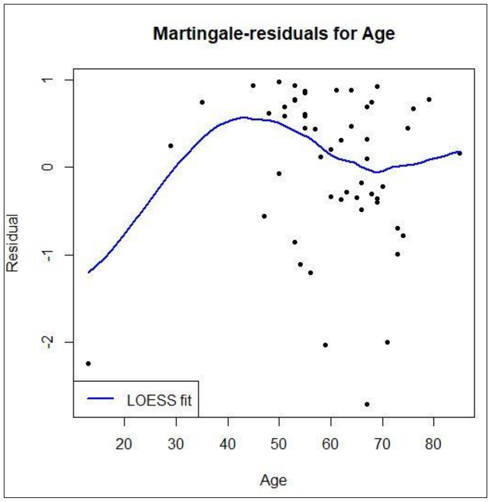


FIGURE 3. Plot of Martingale residual of the model against Age

TABLE 5. The results of Schoenfeld residuals test.

Models	Model 1			Model 2		
Variables	rho	chisq	*P*-value	rho	chisq	*P*-value
Age	0.363	5.204	0.023	0.265	2.969	0.085
Age^2	-	-	-	-0.088	0.322	0.570
Gender	-0.294	4.683	0.031	-0.344	5.912	0.015
Races	0.194	1.711	0.191	0.213	0.870	0.171
Type	-0.092	0.288	0.592	0.008	0.002	0.963
Treatment	-0.300	2.127	0.145	-0.347	3.230	0.072
GLOBAL	NA	10.345	0.066	NA	10.20	0.117

Meanwhile, using the individual or separate test at the 0.05 level of significance, the Age in Model 1 failed the individual test, while the Age and Age2 in Model 2 passed the proportionality test. However, the variable Gender appears to violate proportionality in both models, with P-values of 0.031 and 0.015, respectively. This means that Gender has the potential to change over time.

As a result, this study fit two models: the linear model (Model 1) and the quadratic model (Model 2). By using the quadratic model for the cox proportional hazard to fix on the Age variable and splitting the Gender variable into three different groups Table 6 summarises Model 2 based on the Cox proportional hazard model to achieve the proportionality assumption.

TABLE 6. The results of cox regression for Model 2.

Variables	Estimate	Exp (Estimate)	Standard error	*z*-value	*P*-value
Age	-0.003	0.997	0.009	-0.295	0.768
Age^2	-0.001	0.999	0.000	-1.784	0.074
Races	-0.216	0.806	0.198	-1.093	0.274
Type	0.222	1.249	0.102	2.182	0.029
Treatment	0.158	1.171	0.171	0.921	0.357
Gender:strata (tgroup=1)	0.197	1.218	0.337	0.584	0.559
Gender:strata (tgroup=2)	-0.004	0.996	0.396	-0.010	0.992
Gender:strata (tgroup=3)	-0.212	0.809	0.465	-0.456	0.649
Likelihood ratio test	12.97	Overall *P*-value		0.1128	

This method is used to assess the model's improvement as the functional form is corrected. If the first model is used in this study, two variables must be checked on the varying effect. Because Model 2 only has one covariate that violates proportionality, and that is the categorical variable of Gender, this study will have to proceed with the splitting method, in which the data was divided into three groups: 3 months, 3-6 months, and greater than 6 months. The results of splitting the data produce 120 observations where Gender was treated as a covariate that varied over survival time.

Except for the Type, none of the covariates were found to be significant in terms of survival time. Using the 0.05 significant level, since P-value = 0.029 = 0.05, types of lung cancer were significant in terms of lung cancer survival data. Nonetheless, males outnumber females, older patients die sooner, Chinese outnumber all other races, and chemoradiotherapy was the most commonly used treatment for lung cancer disease.

Assuming, that the proportional hazard assumption holds for models 2 with higher order of age, the results of the Schoenfeld residuals plot from Model 2 are shown in Fig. 4. If a covariate is proportional, the correlations should be close to zero. Large correlation values indicate that the covariates are not proportional. As a result, the graph shows no obvious evidence or pattern that contradicts the proportional hazard assumption.

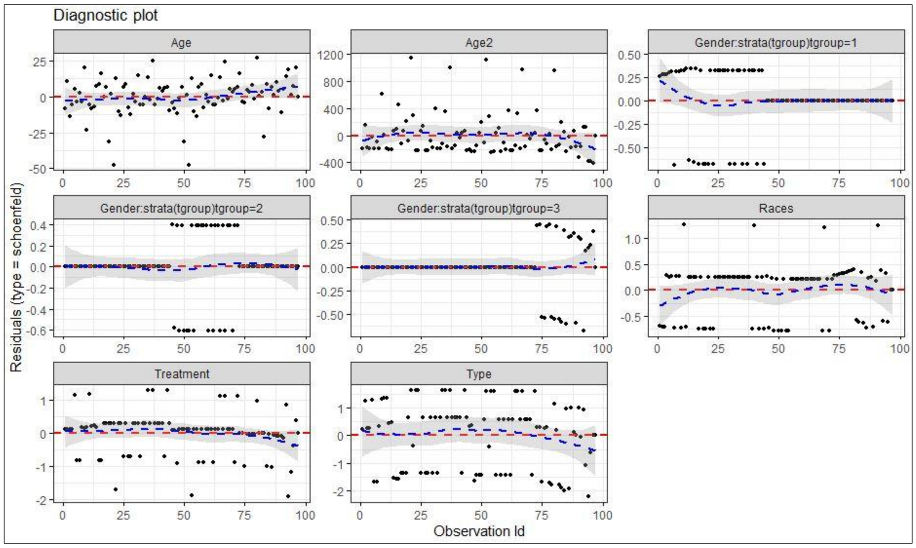
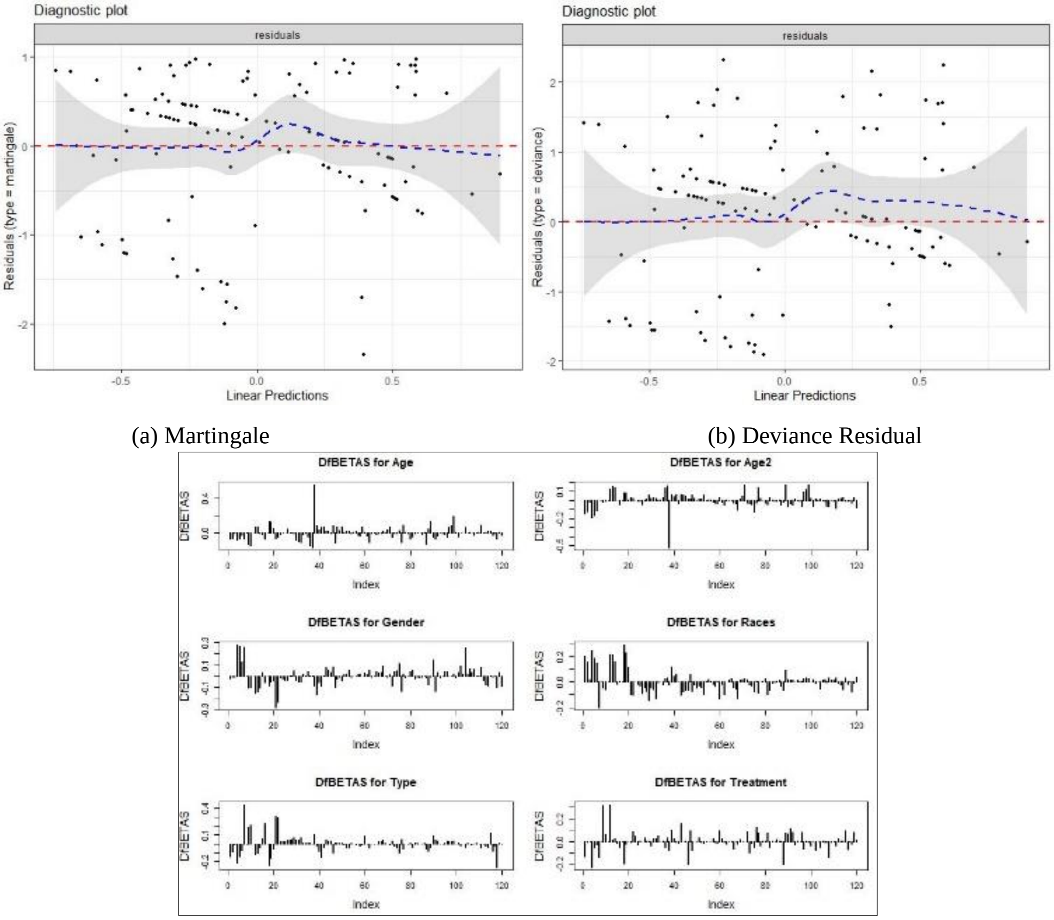


FIGURE 4. Schoenfeld residuals for each covariate against the survival time



(a) Martingale

(b) Deviance Residual

(c) *Dfbetas* residuals

FIGURE 5. Martingale, deviance residual and *Dfbetas* residuals for covariates of Model 2

Furthermore, the martingale residual plot shown in Fig. 5 below was useful in visualising the dependence of lung cancer data where some observations may be censored over survival time. Continuously, in terms of outlying observations, a transformation of the martingale residual leads to the deviance residual, where the deviance will be approximately normal distribution when the censoring rate is less than 25%, while when the censoring rate is greater than 40%, most of the observations will lie within zero values, causing the distribution to be non-normal.

As a result, the following figure is used to discover potential observations on parameter estimation by plotting the dfbetas residuals. The dfbetas residual has been constructed on the left side of Fig. 5 to observe the influential observation for all covariates involved in Model 2. According to the findings of this study, there were no significant influential observations on the lung cancer data.

Parametric Assumption

This study used a normal probability plot to test the parametric assumption, which takes the continuous dependent variable of survival time using the quantile-quantile plot. The plot of normal probability is shown in the figure below.

This study used a normal probability plot to test the parametric assumption, which takes the continuous dependent variable of survival time using the quantile-quantile plot. The plot of normal probability is shown in the figure below.

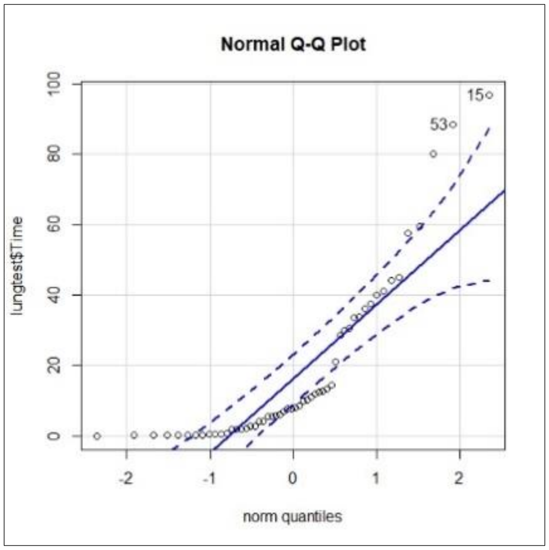
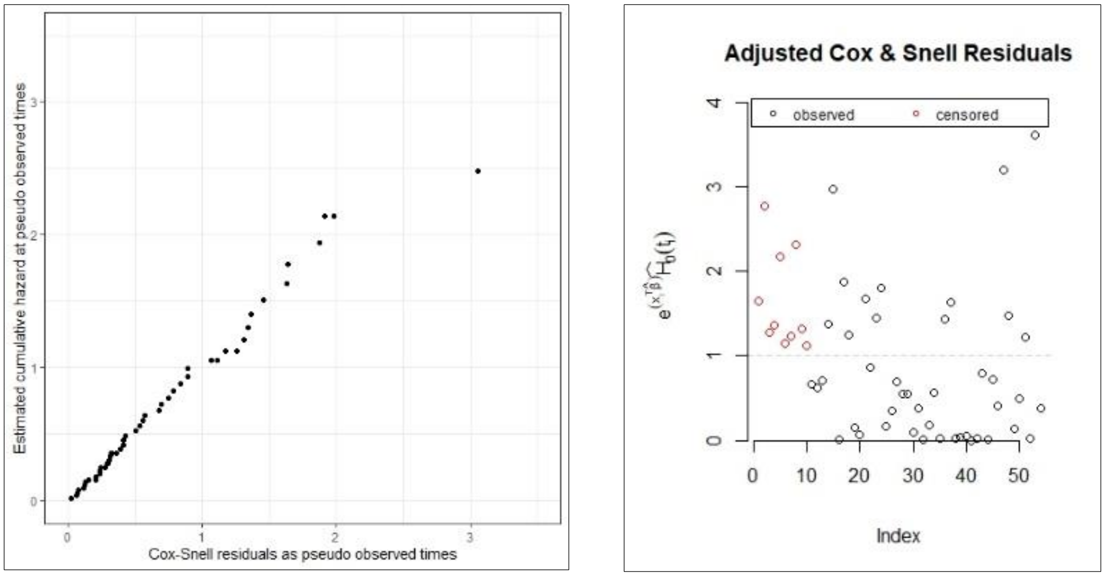


FIGURE 6. Quantile- quantile plot of survival time

According to Fig. 6, the data points fall roughly along a straight line. However, some of the data points from the 53rd and 15th data points deviate from the straight line systematically.

Furthermore, based on Fig. 7, the Cox-Snell residuals show that the survival times follow an exponential distribution with a 45o diagonal line. However, it appears that there is one obvious outlier in the graph plotting. As a result, adjusted Cox-Snell residuals show the adjustment of the survival observation where the censored observation was included to adjust the residual values.



(a) Cox-Snell residuals

(b) Adjusted Cox-Snell residuals

FIGURE 7. Cox-Snell residuals and Adjusted Cox-Snell residuals

CONCLUSION

In short, the analysis started with checking on the covariates of continuous and categorical data. This study found that, the high order of model was needed for the variable of age to be significant. Then, quadratic form of model which was Model 2 has been developed. Additionally, the variable of gender appears to be varying over time since the test of varying effect showed the violation of proportional hazard for variable of gender. Continuously, split method had been applied to deal with the varying effect of gender and the final model has been evaluated. The graphical methods of all categorical variables with regards to its survival time had been shown for martingale residual (form of covariates), *dfbetas* residuals (influential observation), deviance residuals (outliers) and the Cox-Snell residuals (overall fit).

To sum up all the assumptions in this section, the following conclusions are being reached:

1. Univariable analysis: the individual test of each covariate appears to be insignificant in terms of lung cancer survival data.
2. Multivariable analysis: the overall model revealed insignificant covariate values, and the individual test also revealed no significant covariates.
3. One of the categorical variables of gender violated the proportional hazard assumption.
4. The violation of proportional hazard demonstrated that the covariates were not fixed in time and was solved using the split method.
5. The parametric assumption was met because the quantile-quantile plot displays a consistent pattern around 45o. Some outliers have also been identified.
6. The overall fit model using Cox-Snell residuals demonstrates that the distribution fits the lung cancer data, and the Cox model is also quite good at dealing with time-varying covariates.

Since the assumptions for semi-parametric cox-proportional hazard and non-parametric Kaplan-Meier analysis have been violated, necessitating the use of parametric survival analysis methods for this lung cancer data.

ACKNOWLEDGEMENT

We would like to thank College of Computing, Informatics and Mathematics, KPPIM, Universiti Teknologi MARA (UiTM), Shah Alam for their support in this research study.

REFERENCES

1. A. Alotaibi, A. Ali, C. Brown, & F. Sherbeny, J Pharm. Innov. **13**, 15-23 (2022).
2. S. Ainomugisha, “Comparison of semi-parametric and parametric survival analysis models for identifying predictors of virological suppression among HIV-infected adults on ART at the Joint Clinical Research Centre, Uganda”. Doctoral dissertation, Makerere University, 2022.
3. A. P. Kurniati, G. A. A. Wisudiawan, & G. P. Kusuma, “Potentials of Clinical Pathway Analysis Using Process Mining on the Indonesia National Health Insurance Data Samples: an Exploratory Data Analysis,” International Conference on Data Science and Its Applications (ICoDSA,2022*)*, pp. 294-299.
4. E. Rossi, I. G. Zizzari, A. Di Filippo, A. Acampora, M. M. Pagliara, M. G. Sammarco and M. Nuti, Hum. Vaccin. Immunother. **18**, 2034377, (2022).
5. C. Nieder, B. Mannsåker, & R. Yobuta, Cureus **14**, (2022).