Obehi Winnifred Ikpea

Renee Reyes

Sanjana Kumar

Survival Analysis On The Lung Cancer Data

STA 4903-001: Applied Survival Analysis

Professor Jerome Keating

25th April 2023

Background

Cancer, a family of diseases characterized by an uncontrolled division of abnormal cells in a part of the body, is the second leading cause of death in the United States. Lung cancer is a type of cancer that starts in the lungs, typically in the cells lining the bronchi and alveoli, and may spread to the lymph nodes and other organs in the body. It is categorized into small-cell and non-small-cell lung cancer, and they occur and are treated differently. The risk factors for lung cancer are smoking, radon, genetics, radiation therapy, and diet, with smoking being the leading cause.

Every year, approximately 8,000 veterans are diagnosed with Lung cancer by Veterans Affairs. About 5,000 veterans die from the illness every year. Veterans have a higher risk of contracting lung cancer, approximately 900,000 yearly, with a lower survival rate than the general population. This is because these patients, mostly older people, typically have a smoking history and were exposed to hazards during military service. Veteran Affairs is an organization that provides patient and federal care to Veterans, and Its Office of Research and Development supports health research. The VA has done numerous studies on lung cancer, such as discovering the connection between sunlight and skin cancer, developing a nicotine patch in 1984, finding the association between delayed surgery and death in patients with early-stage non-small cell lung cancer, and many more.

Data

The data comes from the Veterans' Administration Lung Cancer study, a randomized trial of two treatment regimens for patients with advanced, inoperable lung cancer. There were 137 patients, of which 128 died during the study, and 9 were censored.

The variables are defined as follows. **Treatment** is a dummy variable that indicates whether the patients received a standard treatment or chemotherapy treatment. **Cell type** refers to the type of lung cancer and is categorized into squamous, small cell, adeno, and large. **Time** refers to the survival time in days. **Status** indicates whether the patients are dead or censored. **Karnofsky score** measures general performance on a scale from 0 to 100. **Months** refers to the months from diagnosis. **Age** describes the age of the patients in years. **Prior therapy** indicates whether the patients received therapy before the study.

Project Goal

This project aims to perform survival analysis on the dataset, which involves comparing survival distributions of the treatment groups and fitting parametric regression and Cox proportional hazard models to identify prognostic factors related to survival time.

Univariate Analysis of the Survival Data

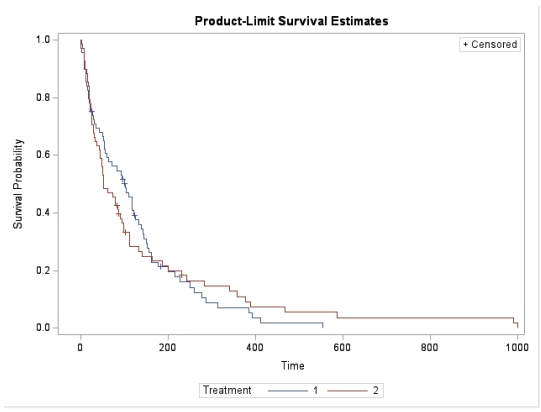
The first analysis conducted was comparing the survival distribution of the treatment group. Patients were randomly assigned to two treatment groups; 69 received the standard treatment, 68 received the chemotherapy treatment, and five were censored from treatment 1 and 4 from treatment 2. We tested the hypothesis

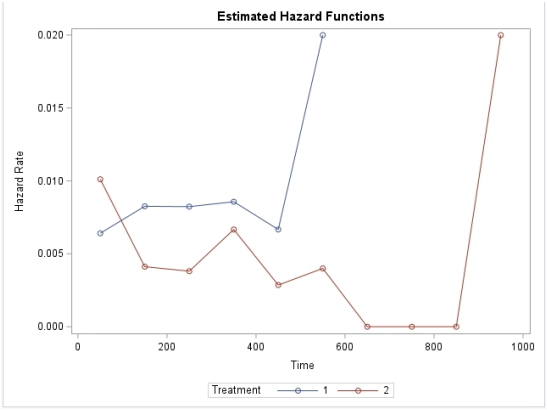
$H_0: S_1(t) = S_2(t)$ $vs.$ $H_1: S_1(t) \neq S_2(t)$ using nonparametric tests like the Log-Rank and Wilcoxon test.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.0082	1	0.9277
Wilcoxon	0.9608	1	0.3270
-2Log(LR)	0.2758	1	0.5995

We failed to reject the null hypothesis at 0.05 level from the results shown above, indicating that there is no sufficient evidence that the two treatments are not equally effective.

We estimated the survival probabilities of the treatment groups using the product limit estimator, a non-parametric method. The survival and hazard curves were obtained, and the following was deduced.





The survival curve showed that the survival distribution for both treatment groups was similar, confirmed by the log-rank test. However, patients who received the chemotherapy treatment had a longer survival time than the latter. The median survival time was also calculated for the treatment groups. The median survival time is the shortest survival time for which the survivor function is less than or equal to 0.5. The median lifetime was estimated to be around 103 days for the standard treatment group and 53 days for the chemotherapy treatment group. We also observed that the hazard curve for treatment 2 - the chemotherapy treatment looked like a bathtub curve, characterized by a mixture of early decreasing and late increasing hazards.

Model

We fitted different parametric and Cox proportional hazards models to investigate the relationship between the survival time and the covariates and identify the risk factors associated with survival time. We first assumed the survival time followed an exponential distribution characterized by a purely random failure pattern and the lack of memory property, and it had a direct relationship with the covariates. Then, using the LIFEREG procedure in SAS, we fitted an exponential model on the survival data. From the output, we saw that the only significant covariates were the Treatment, cell type, and Score, as their Chi-square value was greater than 1. We also observed that CellType 3 was the most significant among the other cell types.

Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	3.3922	0.4639	2.4830	4.3015	53.47	<.0001
Treatment		1	-0.2012	0.1932	-0.5798	0.1774	1.08	0.2976
CellType	1	1	0.3696	0.2719	-0.1634	0.9026	1.85	0.1741
CellType	2	1	-0.4240	0.2580	-0.9298	0.0818	2.70	0.1004
CellType	3	1	-0.7123	0.2905	-1.2816	-0.1430	6.01	0.0142
CellType	4	0	0.0000	.	.	.	.	.
Score		1	0.0297	0.0048	0.0202	0.0391	37.66	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		
Weibull Shape		0	1.0000	0.0000	1.0000	1.0000		

From the table above, we computed the log-survival time model as:

$$log(T) = 3.3922 - 0.2012Treatment + 0.3696CellType_1 - 0.4240CellType_2 - 0.7123CellType_3 + 0.0297Score$$

The log hazard model as:

$$log\, h(t) = -3.3922 + 0.2012Treatment - 0.3696CellType_1 + 0.4240CellType_2 + 0.7123CellType - 0.0297Score$$

- $e^{0.2012} = 1.22$. This indicates that keeping the other covariates constant, the expected failure time for those who received the standard treatment is 22 percent more than for those who received the chemotherapy treatment. This can also be interpreted as patients who recieved the standard treatment are dying in a rate estimated to be 1.22 times that of patients with chemotherapy treatment.
- $e^{-0.3696} = 0.691$. This indicates that controlling for other covariates, the expected failure time for those with squamous cell lung cancer is 69 percent less than for those with large cell lung cancer.
- $e^{0.4240} = 1.52$. This indicates that controlling for other covariates, the expected failure time for those with squamous cell lung cancer is 52 percent more than for those with large cell lung cancer.
- $e^{0.7123} = 2.03$ indicates that controlling for other covariates, the expected failure time for those with adenocarcinoma cell lung cancer is 2.03 percent more than for those with large cell lung cancer.
- $100(e^{-0.0297} - 1) = -2.93$ indicates that each additional increase in the Karnofsky score is associated with a 2.93 percent decrease in the expected failure time, holding other covariates constant.

Since the distribution of the survival time was unknown, and our parametric model was based on assumptions, we decided to fit a Cox proportional model on the survival data. The cox proportional hazard model does not require knowledge of the underlying distribution.

Using PROC PHREG we applied the stepwise procedure to pick out the significant covariates for our proportional hazards model which were Cell and KPS. From our output, the hazard ratio is already estimated for us.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
CELL	1	1	-0.32515	0.27669	1.3809	0.2400	0.722	CELL 1
CELL	2	1	0.38700	0.26100	2.1985	0.1381	1.473	CELL 2
CELL	3	1	0.82566	0.29333	7.9230	0.0049	2.283	CELL 3
KPS		1	-0.03090	0.00518	35.6122	<.0001	0.970	

Model

- h(t|x) = h0(t) * exp(-0.32512 * cell1 + 0.38700 * cell2 + 0.82566 * cell3 - 0.03090 * KPS)

Estimated Hazard Ratios:

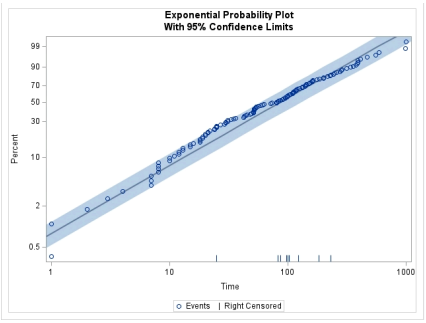
- The hazard ratio of CellType 2 compared to CellType 1 is: $e^{(0.38700)} = 1.473$. Indicating that patients with CellType 2 have a 47% higher hazard of death compared to patients with CellType 1, holding all other covariates constant.
- The hazard ratio of CellType 3 compared to CellType 1 is: $e^{(0.82566)} = 2.283$. Indicating that patients with CellType 3 have a 128% higher hazard of death compared to patients with CellType 1, holding all covariates constant.
- The hazard ratio of KPS score is: $e^{(-0.03090)} = 0.97$. Indicating that for each unit increase in KPS score, the hazard of death decreases by 3%, holding all covariates constant.

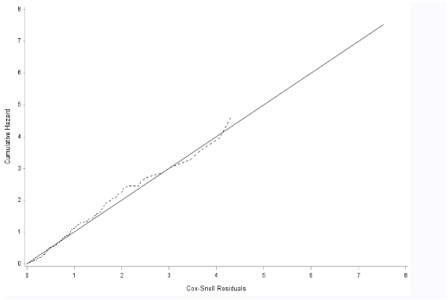
As seen in our exponential model, it should be mentioned that Celltype 3 is the most significant in comparison to the others.

Diagnostics

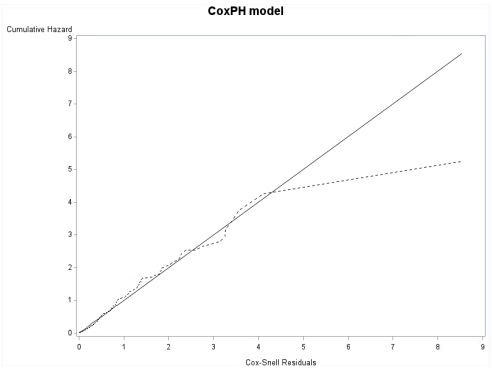
Different measures were implemented to assess model fit. These include the model fit statistics, probability plot, Cox Snell residual, martingale residual, and deviance residual plots.

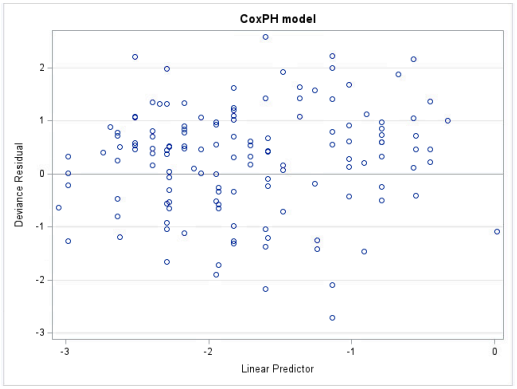
Fit Statistics	
-2 Log Likelihood	394.035
AIC (smaller is better)	406.035
AICC (smaller is better)	406.682
BIC (smaller is better)	423.555

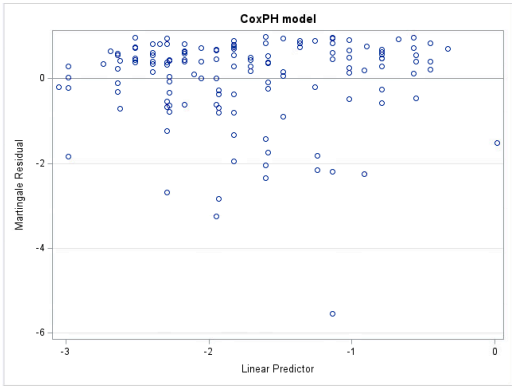




Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1011.768	952.997
AIC	1011.768	960.997
SBC	1011.768	972.405







Model Fit Diagnostics

- From the Model Fit Statistics table, we see that the AIC value for the exponential model is smaller than that for the cox proportional model.
- The probability plot shows that the exponential model fits the data well.
- In the Cox Snell Residuals, we do see that it deviates from the line. Indicating that the Cox Proportional Hazard Model may be unfit for this data. With this we can also see that there is some unexplained variation that the model is not accounting for. We can do a cross validation with the residuals to further investigate if it is or isn't a fit.
- The Martingale residuals show a concern that the model may not be a good fit since a slight U- shape pattern can be seen. Also in the martingale residuals, we can see that there are a couple points raising concern at the bottom of the graph. Using the Deviance residuals as our cross validation, we see the points of concern are points to be concerned about.

From the residuals, above we can see that Cox Proportional Hazard Model is not a fit due to the Cox Snell Residuals strong deviation from the line and the "abnormal" points in the Martingale Residuals. Also the AIC, SBC are higher for this model. All indicating that the exponential model is the better choice for this data.

Summary

Lung cancer is the leading cause of cancer death in the United States, accounting for 1 in 5 deaths. It is prevalent among veterans who have had access to environmental hazards during the military, are old, and have a smoking history. The Veteran Affairs Office of Research and Development conducts research to improve veterans' health. One research involved a Lung cancer trial where patients were randomly given either

standard or chemotherapy treatment, and different variables were measured in the study. We first compared the survival distribution of the two treatment groups and discovered that the treatments were equally effective. We also observed that the median survival time for patients in the standard treatment group was longer than those in the chemotherapy treatment group. To understand how the covariates influence the survival time of the patients, we fitted an exponential and Cox proportional hazard model to the survival time, and from our model diagnostics observation, we concluded that the exponential model fits the data better.

APPENDIX

A. The median survival time:

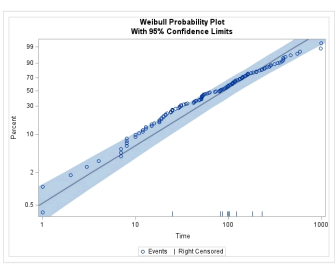
- **Treatment 1-Standard treatment**

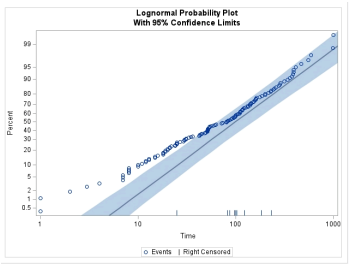
Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	162.000	LOGLOG	132.000	250.000
50	103.000	LOGLOG	54.000	126.000
25	27.000	LOGLOG	12.000	54.000

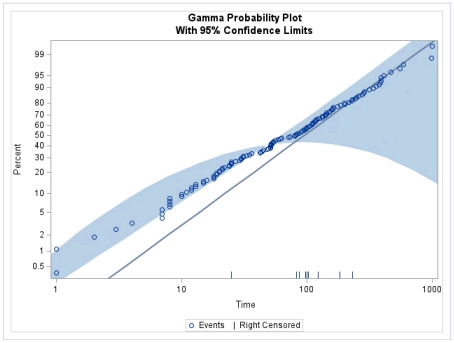
- **Treatment 2-Chemotherapy treatment**

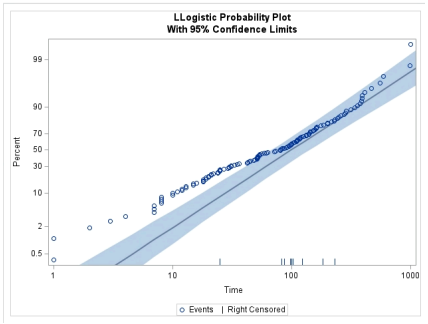
Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	140.000	LOGLOG	99.000	283.000
50	52.500	LOGLOG	43.000	90.000
25	24.500	LOGLOG	15.000	33.000

B. Probability plot of other parametric models attempted:









References

https://www.research.va.gov/programs/pop/lpop.cfm#:~:text=Oncology%20Program%20(LPOP)-,About,survival%20than%20the%20general%20population

https://www.stat.rice.edu/~sneeley/STAT553/Datasets/survivaldata.txt

https://www.cancer.org/cancer/lung-cancer/about/what-is.html

Lee, E. T., & Wang, J. W. (2013). *Statistical methods for survival data analysis*. Wiley.