Project

AS9004A

Survival Analysis

Survival Analysis of Lung Dataset

Author:	*Student number:*
Aryan Rezanezhad	251386495
Jiachen Pan	251023256

Tuesday 12th December, 2023

1 Introduction

Lung cancer has always been one of the most serious cancers in the world, especially in the advanced stages, where survival rates and treatment outcomes are of great concern [1]. Despite recent advances in the detection, pathologic diagnosis and treatment of lung cancer, many patients still develop advanced stages that are incurable and progressively fatal.[2]

Survival analysis methods provide a powerful tool when conducting an analysis of survival time in patients with advanced lung cancer. The method of survival analysis can help us to understand the temporal pattern of disease progression and the factors that may have an impact on the survival time of the patient. Such methods are particularly useful for dealing with events that may not have occurred by the end of the study, a common problem in clinical trials and medical research.

”Right Censor” data are characterized by the fact that the survival time of certain individuals is unknown at the end of the study. This type of data requires the use of special statistical techniques, such as Kaplan-Meier estimators or Cox proportional risk models, to deal with this type of censoring. These methods can provide reasonable estimates of the distribution of survival times as well as interpretability of associated factors.

The purpose of our study was to investigate the key variables that influence the distribution of survival time in patients with advanced lung cancer. In particular, we hoped to answer the following core research question:

- What factors may significantly influence the distribution of survival time?
- What is the distribution of survival time (T)?
- For important parameters, How to find point estimates and interval estimates.

The following report will be structured as follows. In section 2, we will briefly describe the lung data set. In section 3, we will describe the main methods that appear in this article. In section 4 to section 6 we will present the main results of the research. Finally, in sections 7 and 8, we will describe the conclusion and discussion.

2 Data Description

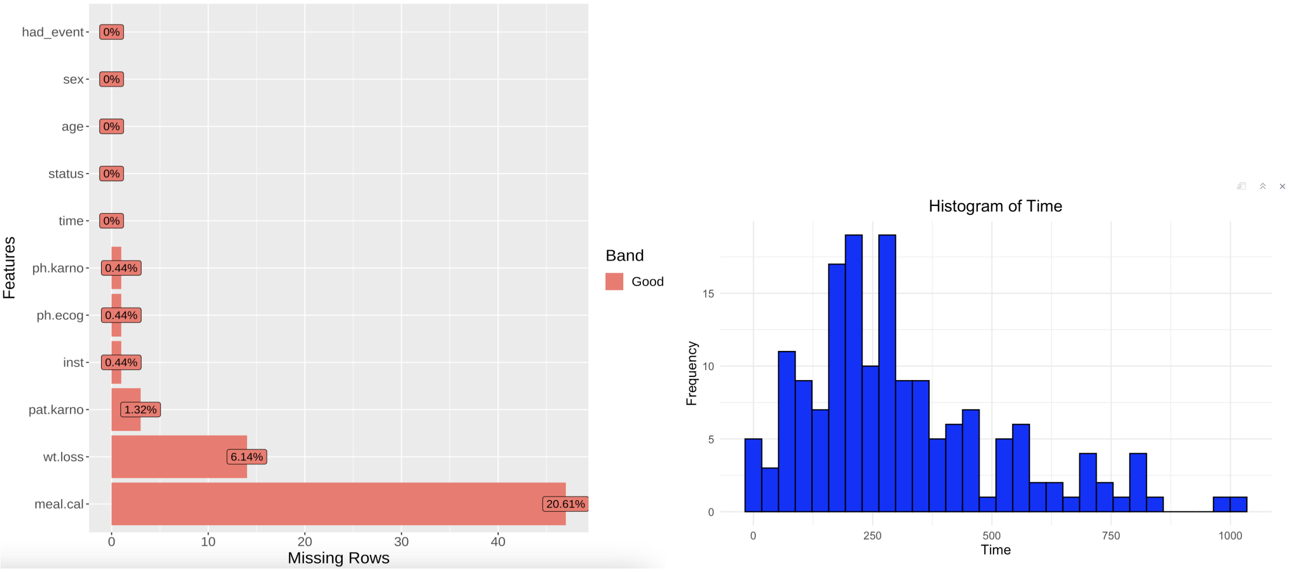


Figure 1

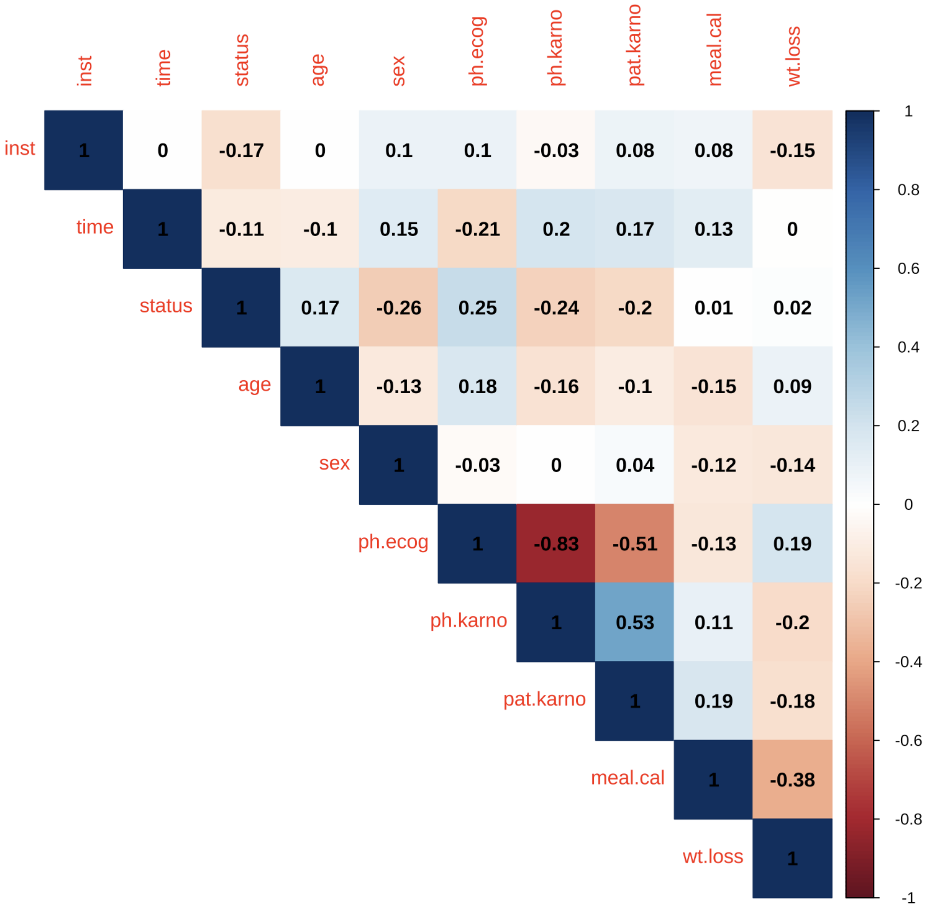
Our dataset comprises 228 observations across nine variables. Excluding the institutional code, the variables are survival time, censoring status, age, ECOG performance score, Karnofsky performance score (assessed both by patients and physicians), calories consumed during main meals, and weight loss experienced by patients in the past. Figure 1 reveals a significant presence of missing values within the dataset, particularly notable in the variable recording calories consumed at main meals. Upon the exclusion of instances with missing values, the dataset retains 167 observations with complete data. Table 1 provides a comprehensive statistical summary of the variables in question. Additionally, second plot on the Figure 1 presents the distribution of the response time, offering a visual representation of the temporal aspect of the data.

Table 1: Summary Statistics of the Advanced Lung Cancer Data Set

Category	Description			Count
status	censored (0)			47
	observed (1)			120
sex	Male (1)			103
	Female (2)			64
ph.ecog	0			47
	1			81
	2			38
	3			1
Continuous	Description	Minimum	Median	Maximum
age	Age of patients in years	39.00	64.00	82.00
ph.karno	Karnofsky performance score (physician)	50.00	80.00	100.00
pat.karno	Karnofsky performance score (patient)	30.00	80.00	100.00
meal.cal	Calories consumed at meals	96.0	975.0	2600.0
wt.loss	Weight loss	-24.000	7.000	68.000

2.1 Correlations

Another aspect of interest in Exploratory Data Analysis is conducting correlation analysis and constructing the correlation matrix to investigate the relationship between the response and the predictors and to check for collinearity among predictors. We apply Pearson's correlation coefficient [3] for numerical variables.



	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
inst	1	0	-0.17	0	0.1	0.1	-0.03	0.08	0.08	-0.15
time		1	-0.11	-0.1	0.15	-0.21	0.2	0.17	0.13	0
status			1	0.17	-0.26	0.25	-0.24	-0.2	0.01	0.02
age				1	-0.13	0.18	-0.16	-0.1	-0.15	0.09
sex					1	-0.03	0	0.04	-0.12	-0.14
ph.ecog						1	-0.83	-0.51	-0.13	0.19
ph.karno							1	0.53	0.11	-0.2
pat.karno								1	0.19	-0.18
meal.cal									1	-0.38
wt.loss										1

Figure 2: Pearson's correlation coefficient

Figure 2 illustrates the high correlation between ECOG performance score and Karnofsky performance score. It is possible that one of the variables may not be interpreted in the model.

3 Methods

The purpose of this section is to introduce some non-parametric methods and parametric models that we will apply to our data.

3.1 Non-parametric methods

3.1.1 The Kaplan-Meier estimator

As a non-parametric statistic, the Kaplan-Meier (KM) estimator [4] is useful for estimating the survival function S(t). It is possible to estimate the true survival function of the population using the KM estimator when the sample size is large enough. KM has the advantage of taking into account some types of censored data, particularly right-censored data. In order to estimate the survival function S(t), the KM estimator is given by the following formula:

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

For index j, t_j represents the time when at least one event occurred, d_j represents the number of events at time t_j, and n_j represents the number of individuals who have survived (have not been censored) until time t_j.

Greenwood's formula [5] can be used to estimate the variance of the KM estimator:

$$\hat{V}ar[\hat{S}_{KM}(t)] = (\hat{S}_{KM}(t))^2 \sum_{j:t_j \leq t} \frac{d_j}{Y_j(Y_j - d_j)} \tag{2}$$

3.1.2 The Nelson-Aalen estimator

Nelson Aalen (NA) estimators [6][7][8], provide yet another non parametric method for estimating the continuous cumulative hazard function. Using the NA estimation of the cumulative hazard function, we can then determine the survival function. Using the same definitions of t_j, d_j and Y_j, for $0 \leq t$,

$$\hat{H}_{NA}(t) = \sum_{j:t_j \leq t} \frac{d_j}{Y_j} \tag{3}$$

and the variance of the Nelson Aalen (NA) estimators:

$$Var[\hat{H}_{NA}(t)] = \sum_{j:t_j \leq t} \frac{d_j}{{Y_j}^2} \tag{4}$$

3.1.3 Conditional Survival Curves

Conditional survival curves provide a dynamic prognosis by recalculating survival probabilities over time. Compared to traditional survival rates, the conditional survival rate reflects the likelihood that a patient will live beyond his or her original prognosis. Let $\hat{S}(t)$ denote the Kaplan-Meier estimator of $S(t)$. Then, $S(t|t_0)$ is consistently estimated by

$$\hat{S}(t|t_0) = \frac{\hat{S}(t + t_0)}{\hat{S}(t_0)} \text{ for } t \geq 0.$$

3.1.4 Log-Rank Test

The log-rank test compares the survival distributions of two samples using a non-parametric hypothesis test. This test evaluates whether there is a statistically significant difference between the groups by comparing the number of events in each group with the number expected from the pooled sample. The null hypothesis is that the two groups have identical hazard functions. The null hypothesis is that the two groups have identical hazard functions, $H_0 : h_1(t) = h_2(t)$. Hence, under H_0, for each group $i = 1, 2$, $O_{i,j}$ follows a hypergeometric distribution with parameters $N_j, N_{i,j}, O_j$. This distribution has expected value

$$E_{i,j} = \frac{O_i N_{i,j}}{N_j}$$

and variance

$$V_{i,j} = E_{i,j} \left(\frac{(N_j - O_j)(N_j - N_{i,j})}{N_j(N_j - 1)} \right).$$

For all $j = 1, \ldots, J$, the logrank statistic compares $O_{i,j}$ to its expectation $E_{i,j}$ under H_0. It is defined as

$$Z_i = \frac{\sum_{j=1}^{J}(O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{for } i = 1 \text{ or } 2$$

3.2 Parametric models

3.2.1 Accelerated Failure Time Model

The Accelerated Failure Time (AFT) model is a parametric survival model which describes the relationship between a set of covariates and the occurrence of an event of interest. The AFT model assumes that covariates accelerate or decelerate the life time of an individual by a constant factor. The model can be represented as $log(T) = X\beta + \sigma W$, where T is the survival time, X is the vector of covariates, β is the vector of coefficients, σ is the scale parameter, and W is the error term, typically assumed to follow a known distribution such as normal, logistic, or Weibull. As a result of this framework, it is possible to interpret the effect of covariates on the time scale directly.

3.2.2 Cox Proportional Hazards Model

The Cox Proportional Hazards Model is a semiparametric model widely used in the field of survival analysis to assess the impact of various covariates on the hazard rate, the instantaneous risk of experiencing the event of interest, such as failure or death, at a particular time. It assumes that covariates have a multiplicative effect on the baseline hazard function, which remains unspecified, allowing the model to focus on the relative effect of covariates without making strong assumptions about the form of the survival function. This model's hazard function can be represented as $h(t|X) = h_0(t)exp(\beta' X)$, where $h(t|X)$ is the hazard at time t given covariates X, $h_0(t)$ is the baseline hazard, β is the vector of coefficients, and $exp(\beta' X)$ is the risk ratio. The proportional hazards assumption implies that the ratio of the hazards for any two individuals is constant over time.

3.2.3 Residual Analysis

In survival analysis, residual analysis is integral for validating model assumptions and fit, particularly in Accelerated Failure Time (AFT) and hazard models. Deviance residuals [9] in AFT models assess model adequacy in capturing time dynamics, indicating fit quality per observation. Conversely, Schoenfeld residuals [10] in Cox hazard models, test the proportional hazards assumption. This analysis is critical as it discerns if covariate effects on the hazard function remain constant over time. These diagnostics are vital for ensuring robustness and reliability in survival model interpretations.

4 Results

4.1 Conditional Survival Curves

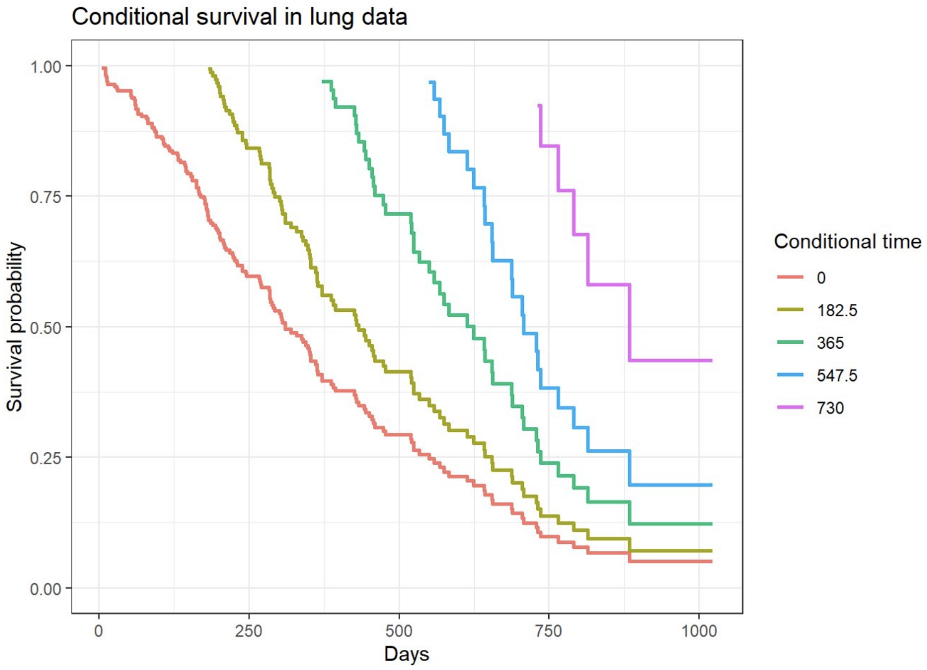


Figure 3: Output of R

We can get estimates and 95% confidence intervals. Let's condition on survival to 6-months. The resulting plot has one survival curve for each time on which we condition. In this case the first line is the overall survival curve since it is conditioning on time 0.

4.2 Survival Probabilities (at 3/6/12/18/24 months)

Months	Estimation	lower	Upper
3	0.88	0.83	0.92
6	0.71	0.64	0.76
12	0.41	0.34	0.48
18	0.26	0.19	0.32
24	0.12	0.07	0.18
6|3	0.8	0.74	0.85
12|3	0.46	0.39	0.54
18|3	0.29	0.22	0.37
24|3	0.13	0.08	0.20
18|12	0.62	0.49	0.73
24|12	0.28	0.17	0.41

Table 2: Codintional and Unconditional Survival Probability

In this table, we can see that for example 0.88 is the probability of one patient who survives up to 3 months and also we can see the lower bound and upper bound of the confidence interval for this patient. Moreover, we can see that the most of patients survives up to 6 months. Also, we can see the conditional survival probability. For instance, 0.8 is the probability of one patient who survives up to 6 months given that be survived in last 3 months. The crucial point to note is that the probability of of survival up to 2 years increased, and it represents the life expectancy plays important role in our life.

4.3 Non-Parametric Methods

In the following Figure, we conducted a comprehensive visual assessment to compare the performance of the Kaplan-Meier (KM) and Nelson-Aalen (NA) methods in modeling the survival function. The results were then meticulously analyzed, revealing a remarkable convergence between the estimates and confidence intervals generated by both methods. This robust similarity indicates that researchers can confidently choose either the KM or NA method to effectively approximate the survival function for their data.

Despite the comparable outcomes, it's essential to delve into the nuances of each method. The Kaplan-Meier method is relatively straightforward and widely used for its simplicity. On the other hand, the Nelson-Aalen method introduces a bit more complexity due to its requirement for an initial estimation of the cumulative hazard function. This additional step can provide a more nuanced understanding of the underlying hazard.

When deciding between these methods, it's crucial to consider the nature of your data and the specific objectives of your analysis. The choice between Kaplan-Meier and Nelson-Aalen should be driven by a thoughtful assessment of the trade-offs between simplicity and a more refined representation of the hazard function.

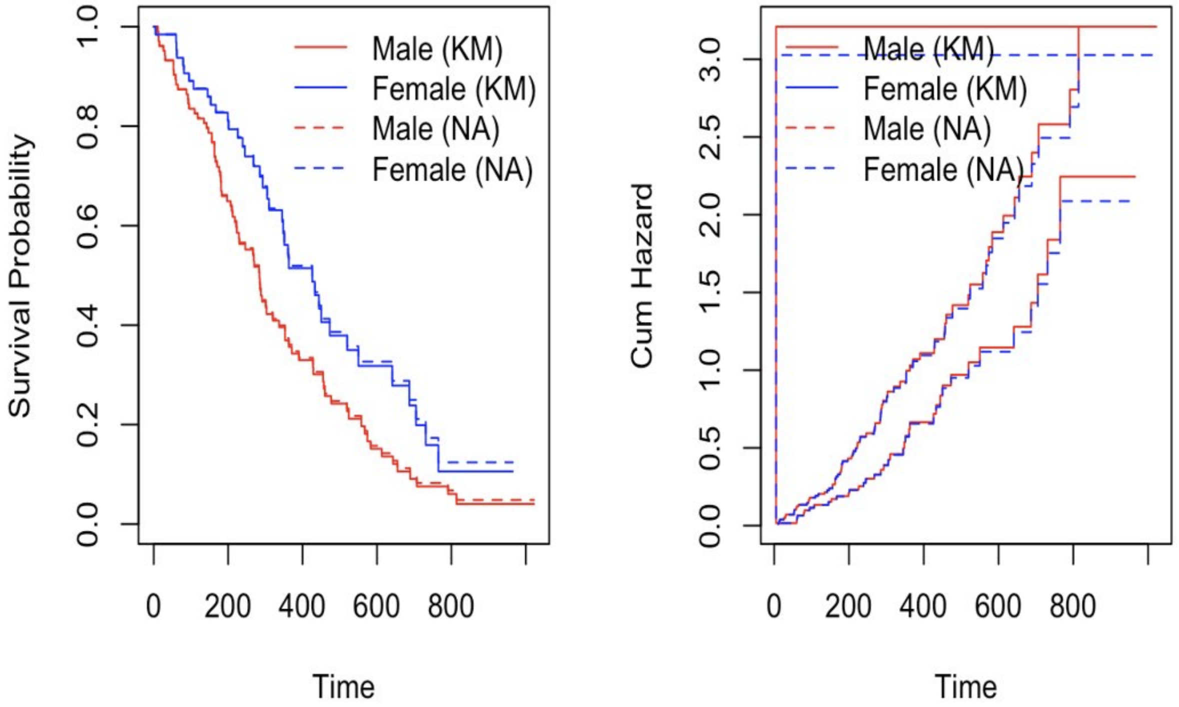


Figure 4: Output of R

4.4 Parametric Methods

In comparing classic parametric models (Weibull, log-normal, and log-logistic) through log(-log(S(t))) vs log(t) plots, the log-normal model emerged as superior. The plots underscore the log-normal model's effectiveness in capturing underlying data patterns. This conclusion aids model selection and suggests the log-normal distribution may better represent observed survival data. Consider these implications for your analysis. Now, let's explore survival distributions for male and female patients. Using the Akaike Information Criterion (AIC) to compare models, the Weibull model exhibited the lowest AIC values in the table. This indicates a preference for the Weibull model over others, suggesting a more effective representation of the observed data and reinforcing its suitability for describing survival patterns.

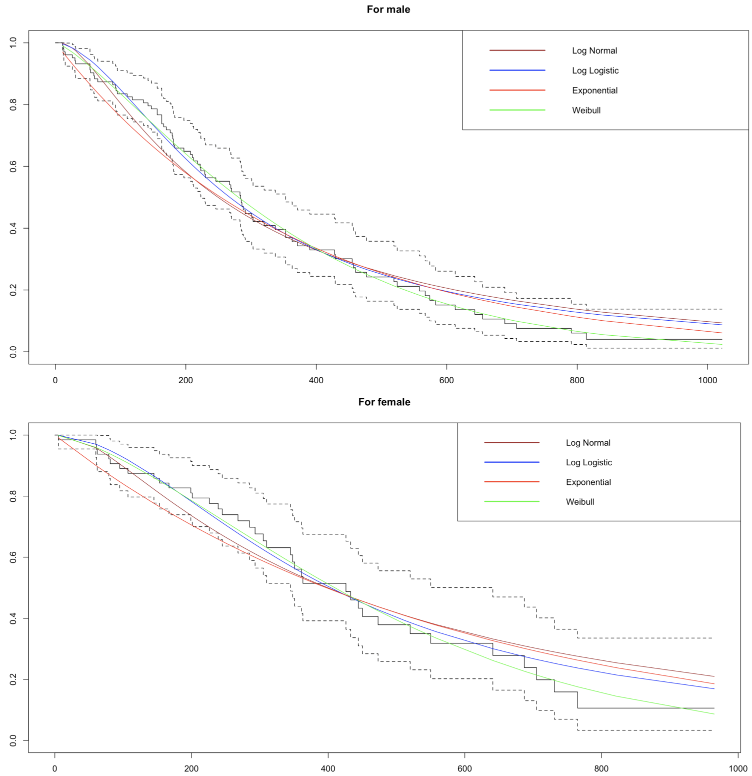


Figure 5: Output of R

Models	**AIC**
Weibull	1127.919
Exponential	1133.901
Log Logistic	1136.218
Log Normal	1142.488

Table 3: AIC Values for Different Models for Male

Models	**AIC**
Weibull	555.407
Exponential	560.6568
Log Logistic	559.1832
Log Normal	565.9405

Table 4: AIC Values for Different Models for female

4.5 Log-Rank Test

The log-rank test can be used to evaluate whether or not Kaplan-Meier (KM) curves for two or more groups are statistically equivalent. The null hypothesis is that there is no difference in survival between the groups.

The log-rank test is a non-parametric test, which makes no assumptions about the survival distributions. It is approximately distributed as a chi-square test statistic.

The function `survdiff()` in the `survival` package can be used to compute the log-rank test comparing males and females' survival curves.

- H_0: There is no difference in the survival function when comparing males to females.
- H_1: There is a difference in the survival function when comparing males to females.

We can see that the P_Value is lower than 0.05. So, we have sufficient evidence to reject the null hypothesis.

Sex	**N**	**Observed**	**Expected**	**P_Value**
1	103	82	68.7	0.01
2	64	38	51.3	

Table 5: Survdiff Test

The log rank test for difference in survival gives a p-value of p = 0.01, indicating that the sex groups differ significantly in survival. There are several alternatives to the log rank test designed to test the hypothesis that two or more survival curves are equivalent called the Wilcoxon, the Tarone-Ware, the Peto, and the Flemington-Harrington test. These test statistics are variations of the log rank test statistic and are derived by applying different weights at the f-th failure time (as shown on the left for two groups).

5 Accelerated Failure Time Model(AFT)

In this phase, we initially constructed a comprehensive model incorporating all relevant explanatory variables.

5.1 Full Model

$$\log(T) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{ph.ecog} + \beta_4 \cdot \text{ph.karno} + \beta_5 \cdot \text{pat.karno} + \beta_6 \cdot \text{meal.cal} + \beta_7 \cdot \text{wt.loss}$$

Variable	**P_Value**
age	0.41613
sex	0.00577
ph.ecog	0.00065
ph.karno	0.03012
pat.karno	0.13389
meal.cal	0.95974
wt.loss	0.07933

Table 6: P_Values for all variables

By looking the results the summary results, we noticed that the p-value for the meal.cal variable was quite large. This means meal.cal didn't seem to have a significant impact on the outcome.

5.2 Remove meal.cal

$$\log(T) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{ph.ecog} + \beta_4 \cdot \text{ph.karno} + \beta_5 \cdot \text{pat.karno} + \beta_6 \cdot \text{wt.loss}$$

Variable	P_Value
age	0.41442
sex	0.00516
ph.ecog	0.00065
ph.karno	0.03009
pat.karno	0.12959
wt.loss	0.07970

Table 7: P_Values for all variables

By looking the results, we noticed that the p-value for the age variable was quite large.

5.3 Remove age

$$\log(T) = \beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog} + \beta_3 \cdot \text{ph.karno} + \beta_4 \cdot \text{pat.karno} + \beta_5 \cdot \text{wt.loss}$$

Variable	P_Value
sex	0.00514
ph.ecog	0.00069
ph.karno	0.04207
pat.karno	0.11858
wt.loss	0.07148

Table 8: P_Values for all variables

By looking the results, we noticed that the p-value for the pat.karno variable was quite large.

5.4 Remove pat.karno

$$\log(T) = \beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog} + \beta_3 \cdot \text{ph.karno} + \beta_4 \cdot \text{wt.loss}$$

Variable	P_Value
sex	0.0041
ph.ecog	5.1e-05
ph.karno	0.0657
wt.loss	0.1293

Table 9: P_Values for all variables

By looking the results, we noticed that the p-value for the wt.loss variable was quite large.

5.5 Remove wt.loss

$$\log(T) = \beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog} + \beta_3 \cdot \text{ph.karno}$$

Variable	P_Value
sex	0.0066
ph.ecog	0.0002
ph.karno	0.0764

Table 10: P_Values for all variables

By looking the results, we noticed that the p-value for the ph.karno variable was quite large.

5.6 Final Model

$$\log(T) = \beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog} = 5.8814 + 0.3727 \cdot \text{sex} - 0.3540 \cdot \text{ph.ecog}$$

Variable	P_Value
sex	0.0098
ph.ecog	0.0003
Models	**AIC**
Full Model	1671.816
Model with sex + ph.ecog	1670.563

By looking at the AFT model, it became evident that the variable sex held statistical significance. The + coefficient, indicates that females generally exhibit a higher probability of survival than males. Also, the ph.ecog is inferred from the - coefficients associated with all ph.ecog factors.

5.7 Residual Plots

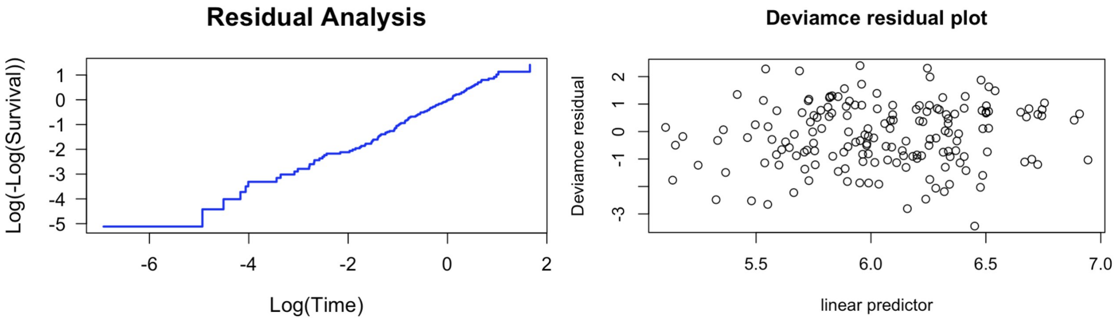


Figure 6: Output of R

From the Deviance Residual plot, we can observe that there is no discernible pattern, indicating that the assumptions of the model are met. The Deviance Residuals appear to be randomly distributed around zero, suggesting that the model adequately captures the underlying patterns in the data.

The absence of a clear trend in the residuals implies that the model's predictions are consistent across the range of observed values. This is an important validation, as a systematic pattern in the residuals could indicate a mis-specification or violation of the model assumptions.

6 Cox Proportional Hazards Model

Similarly to the Accelerated Failure Time (AFT) model, our initial approach involved constructing a Proportional Hazards (PH) model incorporating all relevant explanatory variables. Subsequently, a second model was developed, focusing exclusively on the remaining variables that demonstrated statistical significance:

$$h(t|x) = h_0(t)\exp(\beta' X)$$

6.1 Full Model

$$\beta' X = \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{ph.ecog} + \beta_4 \cdot \text{ph.karno} + \beta_5 \cdot \text{pat.karno} + \beta_6 \cdot \text{meal.cal} + \beta_7 \cdot \text{wt.loss}$$

Variable	**P_Value**
age	0.35168
sex	0.00603
ph.ecog	0.00101
ph.karno	0.04575
pat.karno	0.13685
meal.cal	0.91298
wt.loss	0.06748

Table 11: P_Values for all variables

By looking the results, we noticed that the p-value for the meal.cal variable was quite large.

6.2 Remove meal.cal

$$\beta' X = \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{ph.ecog} + \beta_4 \cdot \text{ph.karno} + \beta_5 \cdot \text{pat.karno} + \beta_6 \cdot \text{wt.loss}$$

Variable	**P_Value**
age	0.35419
sex	0.00531
ph.ecog	0.00102
ph.karno	0.04590
pat.karno	0.13522
wt.loss	0.06808

Table 12: P_Values for all variables

By looking the results, we noticed that the p-value for the age variable was quite large.

6.3 Remove age

$$\beta' X = \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog} + \beta_3 \cdot \text{ph.karno} + \beta_4 \cdot \text{pat.karno} + \beta_5 \cdot \text{wt.loss}$$

Variable	**P_Value**
sex	0.00508
ph.ecog	0.00110
ph.karno	0.06604
pat.karno	0.12008
wt.loss	0.05957

Table 13: P_Values for all variables

By looking the results, we noticed that the p-value for the pat.karno variable was quite large.

6.4 Remove pat.karno

$$\beta' X = \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog} + \beta_3 \cdot \text{ph.karno} + \beta_4 \cdot \text{wt.loss}$$

Variable	P_Value
sex	0.004091
ph.ecog	0.000112
ph.karno	0.100584
wt.loss	0.107975

Table 14: P_Values for all variables

By looking the results, we noticed that the p-value for the wt.loss variable was quite large.

6.5 Remove wt.loss

$$\beta' X = \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog} + \beta_3 \cdot \text{ph.karno}$$

Variable	P_Value
sex	0.007085
ph.ecog	0.000405
ph.karno	0.110009

Table 15: P_Values for all variables

By looking the results, we noticed that the p-value for the ph.karno variable was quite large.

6.6 Final Model

$$\beta' X = \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{ph.ecog} = -0.5101 \cdot \text{sex} + 0.4825 \cdot \text{ph.ecog}$$

Examining the outcomes of the Cox Proportional Hazards Model, it became evident that the variable sex held statistical significance. This implies a notable distinction in survival times between males and females.Also, the AIC value of this new model is smaller than that of the full model.

Variable	P_Value
sex	0.009579
ph.ecog	0.000266
Models	**AIC**
Full Model	1002.069
Model with sex and ph.ecog	1000.751

Table 16: P_Values for all variables

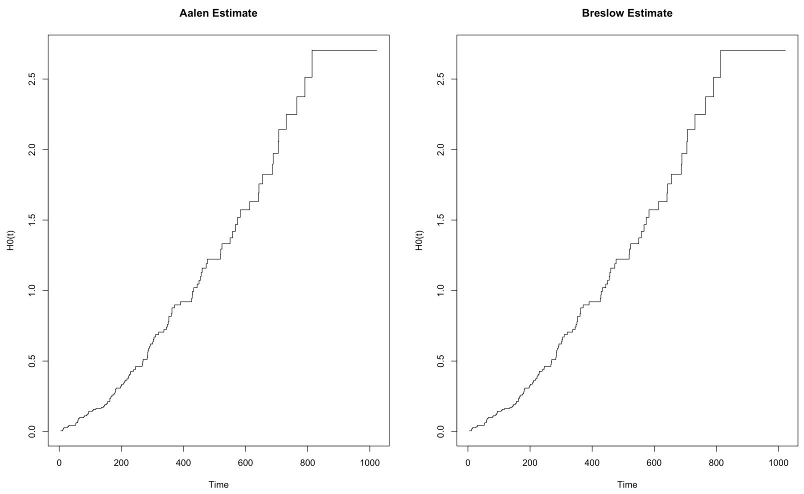


Figure 7: Output of R

These visualizations depict the estimation of the baseline cumulative hazard function H0(t). This outcome enables us to compute the survival function based on the obtained results.

6.7 Validity of the Cox PH model

The Cox proportional hazards model makes several assumptions. We use residuals methods to:

- **Check the proportional hazards assumption** with the **Schoenfeld residuals**.

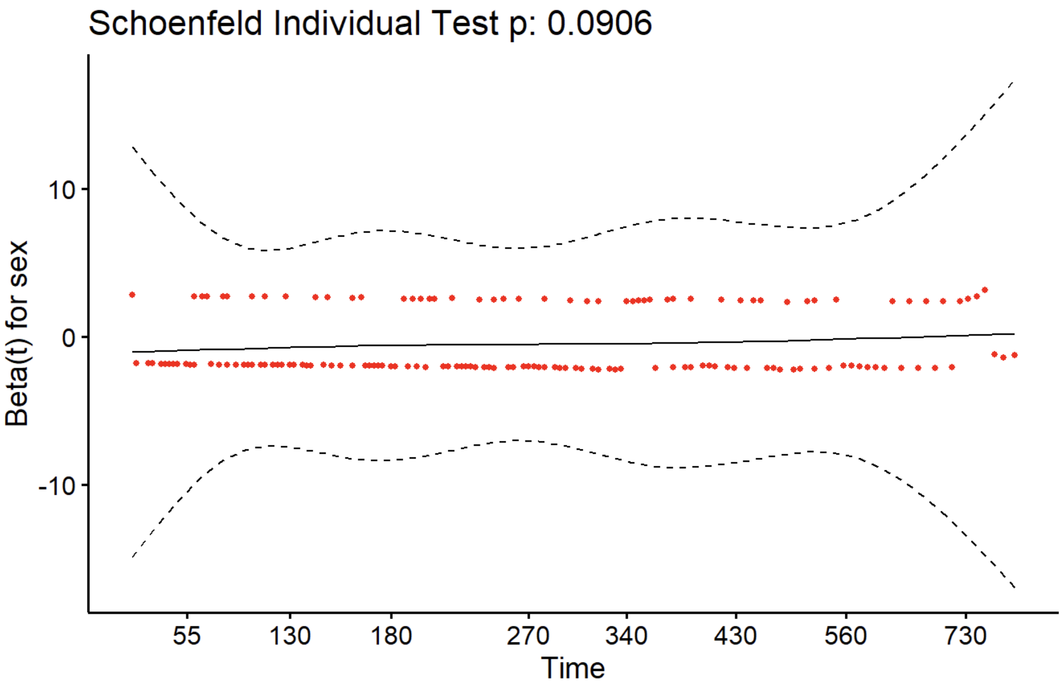


Figure 8: Output of R

From the graphical inspection, there is no pattern with time.

- **Examining influential observations** (or outliers) with **deviance residual** (symmetric transformation of the martinguale residuals), to examine influential observations

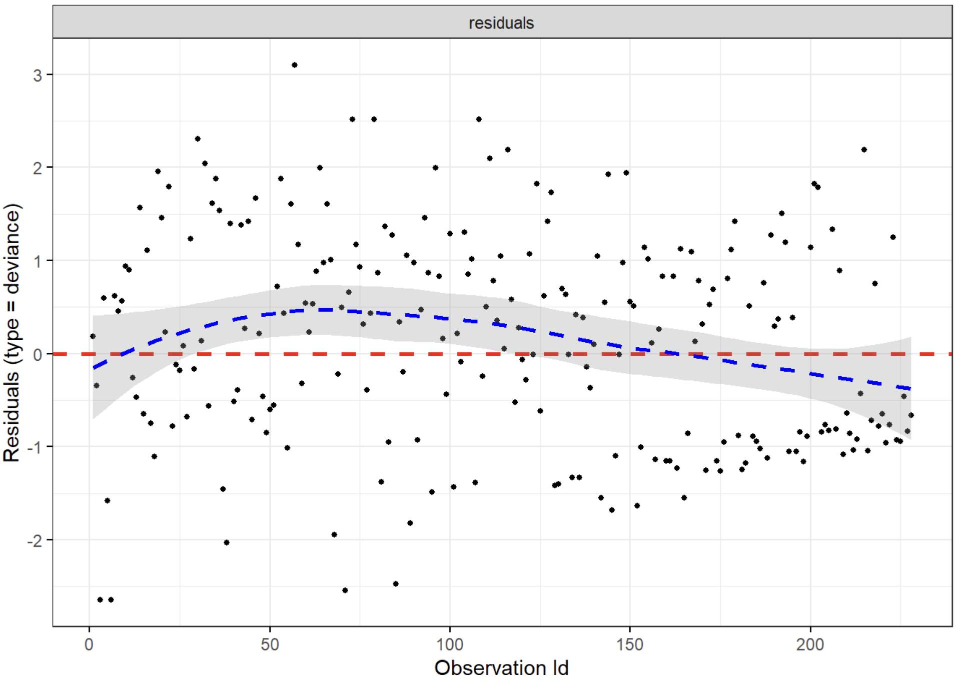


Figure 9: Output of R

These residuals should be roughly symmetrically distributed about zero with a standard deviation of 1.The pattern looks fairly symmetric around 0.

7 Conclusion

Examining the lung cancer dataset, we conducted a thorough assessment by applying both non-parametric and parametric methods and subsequently comparing their efficacy. In the realm of non-parametric approaches, both the Kaplan-Meier (KM) estimator and the Nelson-Aalen (NA) estimator demonstrated commendable performance in terms of estimation accuracy. Transitioning to parametric models, the weibull model exhibited superior performance compared to other models, as indicated by the Akaike Information Criterion (AIC). Also, we employed Accelerated Failure Time (AFT) and Cox Proportional Hazards (PH) models, subjecting them to a comparative analysis. Surprisingly, the AFT model surpassed the PH model in our dataset. While the PH model is more commonly employed, its reliance on the assumption that the hazard ratio remains constant over time can be a limitation. In cases where this assumption is not met, the AFT model may exhibit superior performance compared to the PH model.

8 Roles

Name	**Contents on Reports**
Aryan Rezanezhad	Results + Cox PH Model + AFT Model + Conclusion
Jiachen Pan	Introduction + Data Description + Methods + Analysis of AFT
Name	**Contents on Presentaition**
Aryan Rezanezhad	Cox PH Model + Log Rank + Survival Probabilities
Jiachen Pan	EDA + KM + NA + AFT Model

Table 17: Describing the role of each member

References

[1] W. H. Organization *et al.*, "International agency for research on cancer," 2019.

[2] R. B. L. Lim, "End-of-life care in patients with advanced lung cancer," *Therapeutic Advances in Respiratory Disease*, vol. 10, no. 5, pp. 455–467, 2016, PMID: 27585597. DOI: `10.1177/1753465816660925`. eprint: `https://doi.org/10.1177/1753465816660925`. [Online]. Available: `https://doi.org/10.1177/1753465816660925`.

[3] K. Pearson, "Correlation coefficient," in *Royal Society Proceedings*, vol. 58, 1895, p. 214.

[4] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[5] M. Greenwood *et al.*, "A report on the natural duration of cancer.," *A Report on the Natural Duration of Cancer.*, no. 33, 1926.

[6] W. Nelson, "Hazard plotting for incomplete failure data," *Journal of Quality Technology*, vol. 1, no. 1, pp. 27–52, 1969.

[7] W. Nelson, "Theory and applications of hazard plotting for censored failure data," *Technometrics*, vol. 14, no. 4, pp. 945–966, 1972.

[8] O. Aalen, "Nonparametric inference for a family of counting processes," *The Annals of Statistics*, pp. 701–726, 1978.

[9] T. M. Therneau, P. M. Grambsch, and T. R. Fleming, "Martingale-based residuals for survival models," *Biometrika*, vol. 77, no. 1, pp. 147–160, 1990.

[10] M. Stepanova and L. Thomas, "Survival analysis methods for personal loan data," *Operations Research*, vol. 50, no. 2, pp. 277–289, 2002.

9 Appendix

```
#DataSet
data(cancer, package = "survival")
data <- lung
head(data, n=10)
data <- data |>
  mutate(sex=factor(sex, levels=c(1,2), labels=c("M", "F")),
         had_event=ifelse(status==1, 0, 1))
plot_missing(data, group=c("Good"=1.0),
        theme_config=list(text = element_text(size = 16)))
lung1 <- na.omit(data)
#We want to find the survival distributions for male and female patients
datM<-lung1[lung1[,5]==1,]
datF<-lung1[lung1[,5]==2,]
#Fitting male and female groups using Weibull model
(fitM.w<-survreg(Surv(time, status)~1, data=datM, dist="weibull"))
(fitF.w<-survreg(Surv(time, status)~1, data=datF, dist="weibull"))
#Fitting male and female groups using Exponential model
fitM.e<-survreg(Surv(time, status)~1, data=datM, dist="exponential")
fitF.e<-survreg(Surv(time, status)~1, data=datF, dist="exponential")
#Nonparametric estimate
NfitM<-survfit(Surv(time, status)~1, data=datM)
NfitF<-survfit(Surv(time, status)~1, data=datF)
par(mfrow=c(1,1))
plot(NfitM, main = "Male")
lines(NfitM$time, exp(-NfitM$time*exp(-fitM.e$coef)), col="red")
lines(NfitM$time, exp(-(NfitM$time*exp(-fitM.w$coef))
        ^(1/fitM.w$scale)), col="green")
legend("topright", legend = c("Exponential", "Weibull")
        , col = c("red", "green"), lwd = 2)
plot(NfitF)
lines(NfitF$time, exp(-NfitF$time*exp(-fitF.e$coef)), col="red")
lines(NfitF$time, exp(-(NfitF$time*exp(-fitF.w$coef))
        ^(1/fitF.w$scale)), col="green")
legend("topright", legend = c("Exponential", "Weibull")
        , col = c("red", "green"), lwd = 2)
NfitM.fl<-survfit(Surv(time, status)~1, data=datM, type="fh2")
NfitF.fl<-survfit(Surv(time, status)~1, data=datF, type="fh2")
par(mfrow=c(2,2))
plot(NfitM.fl$time, -log(NfitM.fl$surv), type="s",
    main = "Male - EXP")
plot(log(NfitM.fl$time), log(-log(NfitM.fl$surv)), type="s",
    main = "Male - Weibull")
plot(NfitF.fl$time, -log(NfitF.fl$surv), type="s",
    main = "Female - EXP")
plot(log(NfitF.fl$time), log(-log(NfitF.fl$surv)), type="s",
    main = "Female - Weibull")

# Formal check if weibull can shrink to exp using likelihood ratio
```

```
2*(fitM.w$logl[1]-fitM.e$logl[1])
2*(fitF.w$logl[1]-fitF.e$logl[1])
fitM.l<-survreg(Surv(time, status)~1, data=datM, dist="loglogistic")
fitM.n<-survreg(Surv(time, status)~1, data=datM, dist="lognormal")
fitM.ll<-survreg(Surv(time, status)~1, data=datM, dist="logistic")
fitF.l<-survreg(Surv(time, status)~1, data=datF, dist="loglogistic")
fitF.n<-survreg(Surv(time, status)~1, data=datF, dist="lognormal")
fitF.ll<-survreg(Surv(time, status)~1, data=datF, dist="logistic")
par(mfrow=c(1,1))
plot(NfitM, main = "Male")
lines(NfitM$time, 1/(1+exp((NfitM$time-fitM.ll$coef)
        /fitM.ll$scale)), col="green")
lines(NfitM$time,1/(1+exp((log(NfitM$time)-fitM.l$coef)
        /fitM.l$scale)), col="blue")
legend("topright", legend = c("Logistic", "Log Logistic"),
        col = c("green", "blue"), lwd = 2)
plot(NfitF, main = "Female")
lines(NfitF$time, 1/(1+exp((NfitF$time-fitF.ll$coef)
        /fitF.ll$scale)), col="green")
lines(NfitF$time,1/(1+exp((log(NfitF$time)-fitF.l$coef)
        /fitF.l$scale)), col="blue")
legend("topright", legend = c("Logistic", "Log Logistic"),
        col = c("green", "blue"), lwd = 2)
# Comparing with the nonparametric estimate
par(mfrow=c(1,1))
plot(NfitM, main = "For male")
lines(NfitM$time,1-pnorm((log(NfitM$time)-fitM.n$coef)
        /fitM.n$scale), col="brown")
lines(NfitM$time,1/(1+exp((log(NfitM$time)-fitM.l$coef)
        /fitM.l$scale)), col="blue")
lines(NfitM$time,exp(-NfitM$time*exp(-fitM.e$coef)), col="red")
lines(NfitM$time, exp(-(NfitM$time*exp(-fitM.w$coef))
        ^(1/fitM.w$scale)), col="green")
legend("topright", legend = c("Log Normal", "Log Logistic",
        "Exponential", "Weibull"), col = c("brown", "blue",
                "red", "green"), lwd = 2)
plot(NfitF, main = "For female")
lines(NfitF$time,1-pnorm((log(NfitF$time)-fitF.n$coef)
        /fitF.n$scale), col="brown")
lines(NfitF$time,1/(1+exp((log(NfitF$time)-fitF.l$coef)
        /fitF.l$scale)), col="blue")
lines(NfitF$time, exp(-NfitF$time*exp(-fitF.e$coef)), col="red")
lines(NfitF$time, exp(-(NfitF$time*exp(-fitF.w$coef))
        ^(1/fitF.w$scale)), col="green")
legend("topright", legend = c("Log Normal", "Log Logistic",
        "Exponential", "Weibull"), col = c("brown", "blue",
                "red", "green"), lwd = 2)
AIC(fitM.w) #Weibull
AIC(fitM.e) #Exponential
AIC(fitM.l) #Log Logistic
AIC(fitM.n) #Log Normal
AIC(fitF.w) #Weibull
AIC(fitF.e) #Exponential
AIC(fitF.l) #Log Logistic
AIC(fitF.n) #Log Normal

##Nonparametrically
(fitD<-survdiff(Surv(time, had_event)~sex, data=lung1))
##Parametrically, assume Weibull distribution
fit.w<-survreg(Surv(time, had_event)~sex, data=lung1,
        dist="weibull")
2*(fitM.w$logl[1]+fitF.w$logl[1]-fit.w$logl[1])
summary(fit.w)

# Non-parametric methods
#KM
fit_KM <- survfit(Surv(time,status) ~ 1, data)
fit_fl <- survfit(Surv(time,status) ~ 1, data, type = "fl")
#NA
par(mfrow=c(1,2))
plot(fit_KM, xlab = "Time", ylab = "Survival Probability", conf.int = F,
      lty = 1, mark.time = F)
lines(fit_fl, lty= 2, conf.int = F)
legend("topright", legend = c("Kaplan-Meier", "fl"),
        lty = c(1, 2), bty = "n")

plot(fit_fl$time, -log(fit_fl$surv), type="s", xlab="time",
     ylab="Cum Hazard")
lines(fit_KM$time, -log(fit_KM$surv), type="s", lty=2)
legend("topleft", legend = c("Kaplan-Meier", "fl"),
        lty = c(1, 2), bty = "n")
#AFT
library(dplyr)
fit1<-survreg(Surv(time, had_event)~age+sex+ph.ecog+
        ph.karno+pat.karno+meal.cal+wt.loss,data = lung1,
                dist = "weibull")
summary(fit1)
fit2<-survreg(Surv(time, had_event)~age+sex+ph.ecog
        +ph.karno+pat.karno+wt.loss,data = lung1,
                dist = "weibull")
summary(fit2)
fit3<-survreg(Surv(time, had_event)~sex+ph.ecog+ph.karno
        +pat.karno+wt.loss,data = lung1,dist = "weibull")
summary(fit3)
fit4<-survreg(Surv(time, had_event)~sex+ph.ecog+ph.karno
        +wt.loss,data = lung1,dist = "weibull")
summary(fit4)
fit5<-survreg(Surv(time, had_event)~sex+ph.ecog+ph.karno
        ,data = lung1,dist = "weibull")
summary(fit5)
fit6<-survreg(Surv(time, had_event)~sex+ph.ecog,
        data = lung1,dist = "weibull")
```

```
summary(fit6)
AIC(fit1)
AIC(fit2)
AIC(fit3)
AIC(fit4)
AIC(fit5)
AIC(fit6)
cs.res<-exp(-fit6$linear.predictor/fit6$scale)*
        (Surv(lung1$time, lung1$had_event)[,1])^
                (1/fit6$scale)
cs.fit<-survfit(Surv(cs.res,lung1$had_event)~1, type="fh2")
plot(cs.fit$time, -log(cs.fit$surv), type="s")
plot(log(cs.fit$time), log(-log(cs.fit$surv)), type="s")
plot(log(cs.fit$time), log(exp(-log(cs.fit$surv))-1),
        type="s")
#Deviance residual plot
par(mfrow=c(1,1))
de.res<-residuals(fit6, type="deviance")
plot(fit6$linear.predictor, de.res)
plot(Surv(lung1$time, lung1$had_event)[,1], de.res)

#The effect of sex
NonFit.trt<-survfit(Surv(time, had_event)~sex, data=lung1)
plot(NonFit.trt,col=c("red","blue"))
plot(log(NonFit.trt$time), log(-log(NonFit.trt$surv)),
        col=c("red","blue"))
#The effect of ph.ecog
NonFit.trt<-survfit(Surv(time, had_event)~ph.ecog, data=lung1)
plot(NonFit.trt,col=c("red","blue","green","brown"))
plot(log(NonFit.trt$time), log(-log(NonFit.trt$surv)),
        col=c("red","blue","green","brown"))
(cox.fit1<-coxph(Surv(time, had_event)~age+sex+ph.ecog+
        ph.karno+pat.karno+meal.cal+wt.loss, data=lung1))
(cox.fit2<-coxph(Surv(time, had_event)~age+sex+ph.ecog
        +ph.karno+pat.karno+wt.loss, data=lung1))
(cox.fit3<-coxph(Surv(time, had_event)~sex+ph.ecog+ph.karno+
        pat.karno+wt.loss, data=lung1))
(cox.fit4<-coxph(Surv(time, had_event)~sex+ph.ecog
        +ph.karno+wt.loss, data=lung1))
(cox.fit5<-coxph(Surv(time, had_event)~sex+ph.ecog+ph.karno
        , data=lung1))
(cox.fit6<-coxph(Surv(time, had_event)~sex+ph.ecog,
        data=lung1))
plot(cox.zph(cox.fit6))
AIC(cox.fit1)
AIC(cox.fit6)

par(mfrow=c(3,2))
plot(lung1$time, residuals(cox.fit6, type="martingale"),xlab="Time",ylab=
        main="Martingale")
lines(lowess(lung1$time, residuals(cox.fit6,
        type="martingale")))

plot(lung1$time, residuals(cox.fit6, type="deviance"),
        xlab="Time",ylab="Residual",main="Deviance")
lines(lowess(lung1$time, residuals(cox.fit6,
        type="deviance")))

plot(lung1$time, residuals(cox.fit6, type="score")[,1],
        xlab="Time",ylab="Residual",main="Score Sex")
lines(lowess(lung1$time, residuals(cox.fit6, type="score")
        [,1]))
plot(lung1$time, residuals(cox.fit6, type="score")[,2],
        xlab="Time",ylab="Residual",main="Score ph.ecog")
lines(lowess(lung1$time, residuals(cox.fit6, type="score")
        [,2]))
plot(lung1$time[lung1$had_event==1], residuals(cox.fit6,
        type="scaledsch")[,1], xlab="Time",ylab="Residual"
        ,main="Scaledsch Sex")
lines(lowess(lung1$time[lung1$had_event==1],
        residuals(cox.fit6, type="scaledsch")[,1]))
plot(lung1$time[lung1$had_event==1], residuals(cox.fit6,
        type="scaledsch")[,1], xlab="Time",ylab="Residual",main="Scaledsc
lines(lowess(lung1$time[lung1$had_event==1],
        residuals(cox.fit6, type="scaledsch")[,1]))
par(mfrow=c(1,2))
cox.n.fit<-survfit(cox.fit6, type="aalen")
cox.n.fit1<-survfit(cox.fit6, type="breslow")
plot(cox.n.fit, main = "aalen")
plot(cox.n.fit1, main = "breslow")
## Baseline cumulative hazard function
plot(cox.n.fit$time, -log(cox.n.fit$surv), main="Aalen
        Estimate", type="s", xlab = "Time", ylab = "H0(t)")
plot(cox.n.fit1$time, -log(cox.n.fit1$surv), main="Breslow
        Estimate", type="s", xlab = "Time", ylab = "H0(t)")
### Influential observation detection
par(mfrow=c(1,1))
plot(residuals(cox.fit6, type="dfbeta"))
```