Survival Analysis of Patients with Lung Cancer

P8108 Final Project: Group 5
Chloe Jian, Hening Cui, Jibei Zheng, Pengchen Wang, Xueqing Huang, Qihang Wu

Dec 10, 2022

1 Introduction

Lung cancer is a disease with a very high prevalence. Prognostic factors provide important information for patients with cancer. A better understanding of patients' prognosis can help in making appropriate therapeutic decisions[1]. Driven by the desire to improve life quality of lung cancer patients, we perform a survival analysis of these patients and analyze factors that affect survival time.

The dataset we use is the lung cancer dataset in 'survival' package in R. The data describes survival of patients with advanced lung cancer from the North Central Cancer Treatment Group, as well as measures of the patients performance assessed either by the physician and by the patients themselves[1]. Our project aims to explore whether factors such as age, sex, and caloric intake, will bring significant differences in the survival rate of patients with advanced lung cancer. The association between both the physician's assessments of performance status as well as the patient's assessment of their own performance status and the survival rate are also evaluated.

Methods we use in this project include exploratory data analysis, non-parametric estimate, hypothesis testing, semi-parametric model and parametric models. Details of those methods are given below.

2 Methods

2.1 Exploratory Data Analysis

The dataset contains a total of 228 patients and 10 variables. A brief description of variables in the dataset is shown below.

- inst: Institution code
- time: Survival time in days
- status: Censoring status(1=censored, 2=dead)
- age: Age in years
- sex: Male=1, Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (from bad=0 to good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient
- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in last six months

Survival endpoint is the death of patients. The type of censoring is right censoring, which means patients left the study before their death. Among 228 patients, 63 of them were right censored and the number of

events was 165. We group the patients by their survival status and provide the descriptive statistics of other variables. Wilcoxon rank sum test, Pearson's Chi-squared test, and Fisher's exact test were used to compare values across group.

Table 1: Patient Characteristics

Variable	**Overall**, N = 228	**Alive**, N = 63	**Death**, N = 165	**p-value**
Survival Time (days)	305 (211)	363 (221)	283 (203)	0.003
Age	62 (9)	60 (10)	63 (9)	0.053
Sex				<0.001
Male	138 (61%)	26 (41%)	112 (68%)	
Female	90 (39%)	37 (59%)	53 (32%)	
ECOG Score				0.003
0	63 (28%)	26 (41%)	37 (23%)	
1	113 (50%)	31 (49%)	82 (50%)	
2	50 (22%)	6 (9.5%)	44 (27%)	
3	1 (0.4%)	0 (0%)	1 (0.6%)	
Missing	1	0	1	
Karnofsky Score(by physician)				0.057
50	6 (2.6%)	1 (1.6%)	5 (3.0%)	
60	19 (8.4%)	3 (4.8%)	16 (9.8%)	
70	32 (14%)	3 (4.8%)	29 (18%)	
80	67 (30%)	20 (32%)	47 (29%)	
90	74 (33%)	25 (40%)	49 (30%)	
100	29 (13%)	11 (17%)	18 (11%)	
Missing	1	0	1	
Karnofsky Score(by patients)				0.043
30	2 (0.9%)	1 (1.6%)	1 (0.6%)	
40	2 (0.9%)	1 (1.6%)	1 (0.6%)	
50	4 (1.8%)	0 (0%)	4 (2.5%)	
60	30 (13%)	3 (4.8%)	27 (17%)	
70	41 (18%)	10 (16%)	31 (19%)	
80	51 (23%)	12 (19%)	39 (24%)	
90	60 (27%)	22 (35%)	38 (23%)	
100	35 (16%)	14 (22%)	21 (13%)	
Missing	3	0	3	
Calories Consumed (kcals)	929 (402)	913 (453)	934 (384)	0.4
Missing	47	16	31	
Weight Loss (pounds)	10 (13)	9 (13)	10 (13)	0.3
Missing	14	1	13	

From the table, we can see that average survival time for censored and dead patients is 363 days and 283 days, respectively. From the p values, we can see that for patients who were alive and dead, the survival time, sex proportion, ECOG performance score and Karnofsky performance score rated by patient are significantly different. However, there are no significant differences in age, Karnofsky performance score rated by physician, calories consumed, and weight loss. (Note that the p values for all continuous variables are obtained from Wilcoxon rank sum tests while Fisher's exact tests for the categorical.)

From the table we can see that there are some missing values in this dataset. For simplicity, we removed those missing data for the following analysis.

2.2 Non-parametric Estimate

Lifetable The lifetable was constructed using standard life table methodology[2]. Table 2 represents the lifetable with the time break of 100 days stratified by sex. The full table was in supplementary material. Based on the lifetable, the 50% survival time of male is between 285 to 286 days versus 433-434 for female. Male has lower survival time than female based on lifetable. The hypothesis need future testify in the following modeling fitting.

	tstart	tstop	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv	se.pdf	se.hazard
0-100	0	100	103	0	103.0	17	1.00000000	0.0016504854	0.001798942	0.00000000	0.0003657782	0.0004345389
100-200	100	200	86	5	83.5	19	0.83495146	0.0018998895	0.002567568	0.03657782	0.0003920163	0.0005841662
200-300	200	300	62	7	58.5	18	0.64496250	0.0019845000	0.003636364	0.04760067	0.0004158393	0.0008428131
300-400	300	400	37	3	35.5	9	0.44651250	0.0011320035	0.002903226	0.05099700	0.0003507137	0.0009574916
400-500	400	500	25	3	23.5	6	0.33331215	0.0008510097	0.002926829	0.05012019	0.0003259763	0.0011820092
500-600	500	600	16	0	16.0	6	0.24821117	0.0009307919	0.004615385	0.04787378	0.0003499672	0.0018333649
600-700	600	700	10	0	10.0	4	0.15513198	0.0006205279	0.005000000	0.04239983	0.0002941465	0.0024206146
700-800	700	800	6	0	6.0	2	0.09307919	0.0003102640	0.004000000	0.03499672	0.0002137673	0.0027712813
800-900	800	900	4	2	3.0	1	0.06205279	0.0002068426	0.004000000	0.02941465	0.0001952848	0.0039191836
900-1000	900	1000	1	0	1.0	0	0.04136853	0.0000000000	0.000000000	0.02587989	NaN	NaN
1000-1100	1000	1100	1	1	0.5	0	0.04136853	0.0000000000	0.000000000	0.02587989	NaN	NaN
1100-1200	1100	1200	0	0	0.0	0	0.04136853	NaN	NaN	0.02587989	NaN	NaN
1200-Inf	1200	Inf	0	0	0.0	0	NaN	NA	NA	NaN	NA	NA

Table 2-1 Lifetable (male)

	tstart	tstop	nsubs	nlost	nrisk	nevent	surv	pdf	hazard	se.surv	se.pdf	se.hazard
0-100	0	100	64	0	64.0	7	1.0000000	0.0010937500	0.0011570248	0.00000000	0.0003901364	0.0004365819
100-200	100	200	57	3	55.5	5	0.8906250	0.0008023649	0.0009433962	0.03901364	0.0003440834	0.0004214300
200-300	200	300	49	12	43.0	7	0.8103885	0.0013192371	0.0017721519	0.04931280	0.0004632460	0.0006671758
300-400	300	400	30	4	28.0	7	0.6784648	0.0016961620	0.0028571429	0.06153038	0.0005761152	0.0010688223
400-500	400	500	19	0	19.0	5	0.5088486	0.0013390753	0.0030303030	0.07219475	0.0005480368	0.0013395469
500-600	500	600	14	4	12.0	2	0.3749411	0.0006249018	0.0018181818	0.07397516	0.0004217940	0.0012803251
600-700	600	700	8	0	8.0	2	0.3124509	0.0007811272	0.0028571429	0.07367033	0.0005125726	0.0019995835
700-800	700	800	6	1	5.5	3	0.2343382	0.0012782082	0.0075000000	0.07308191	0.0006375364	0.0040141352
800-900	800	900	2	1	1.5	0	0.1065174	0.0000000000	0.0000000000	0.05982461	NaN	NaN
900-1000	900	1000	1	1	0.5	0	0.1065174	0.0000000000	0.0000000000	0.05982461	NaN	NaN
1000-1100	1000	1100	0	0	0.0	0	0.1065174	NaN	NaN	0.05982461	NaN	NaN
1100-1200	1100	1200	0	0	0.0	0	NaN	NaN	NaN	NaN	NaN	NaN
1200-Inf	1200	Inf	0	0	0.0	0	NaN	NA	NA	NaN	NA	NA

Table 2-2 Lifetable (female)

The Kaplan-Meier and Fleming-Harrington model For nonparametric estimator, Kaplan-Meier (KM) model and Fleming-Harrington (FH) model were used to measure the fraction of subjects living for a certain amount of time after treatment with the stratify of sex[3].

- The Kaplan-Meier estimator

$$\hat{S}_K(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t}[1 - \frac{d_i}{n_i}] & \text{if } t \geq t_1 \end{cases}$$

note: $d_i = \#\ of\ failure\ at\ time\ t_i$, $n_i = \#\ at\ risk\ at\ t_i^-$, $c_i = \#\ censored\ during\ the\ interval\ [t_i, t_{i+1}]$

- The Fleming-Harrington estimator

$$\hat{S}_F(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} exp[-\frac{d_i}{n_i}] & \text{if } t \geq t_1 \end{cases}$$

Both the KM estimator and the FH estimator have P-values that are lower than 0.05. We have 95% confidence that there are differences between the survival curves over the male and female. According to the following graph(Figure 1), male have a lesser chance of living through the 3 years than female. The difference between sex is more significant in the early time point than the later time point.

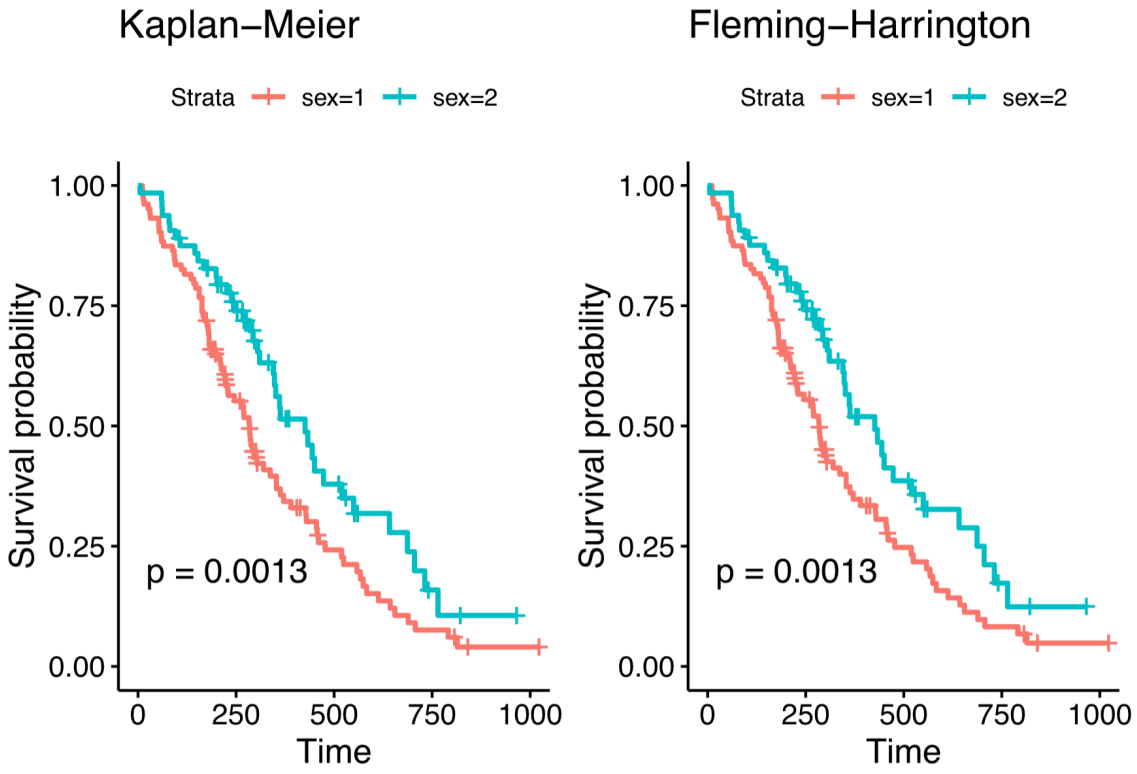

Figure 1: Kaplan-Meier model and Fleming -Harrington model (sex=1 (male), sex=2 (female))

The KM and FH model have similar trend with no significant difference. FH has slight higher estimator in the late time point than KM estimator.

2.3 Hypothesis Testing

Log-Rank test and Wilcoxon test Non-parametric test was conducted to compare the survival experience between males and females, using Log-Rank test and Wilcoxon test (Table 3). However, according to the SAS output, inconsistent results were generated. The Log-Rank test gave us a test statistic of 6.23 with a p-value of 0.01. Thus, we will reject the null hypothesis and conclude that there is a significant difference in survival experience between males and females. However, when we move on to use the Wilcoxon test, chi-square statistic is 3.04, with its associated p-value 0.08. We thus fail to reject the null hypothesis and conclude that there is no significant difference between two sex groups. The underlying reason behind this discrepancy is due to the fact that the Wilcoxon test is more sensitive to detect early points in time than the Log-Rank test.

Survival Curve By looking at survival curves for males and females (Figure 2), apparently in the earlier point in time, there is no obvious difference in survival probability between males and females. Nevertheless, for the later point in time,the gap starts getting larger. This gap explains why we got a non-significant result from the Wilcoxon test: it has more weights to detect earlier time that has no significant difference.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	6.2289	1	0.0126
Wilcoxon	3.0413	1	0.0812
-2Log(LR)	5.6211	1	0.0177

Table 3: Comparison of survival experience between males and females

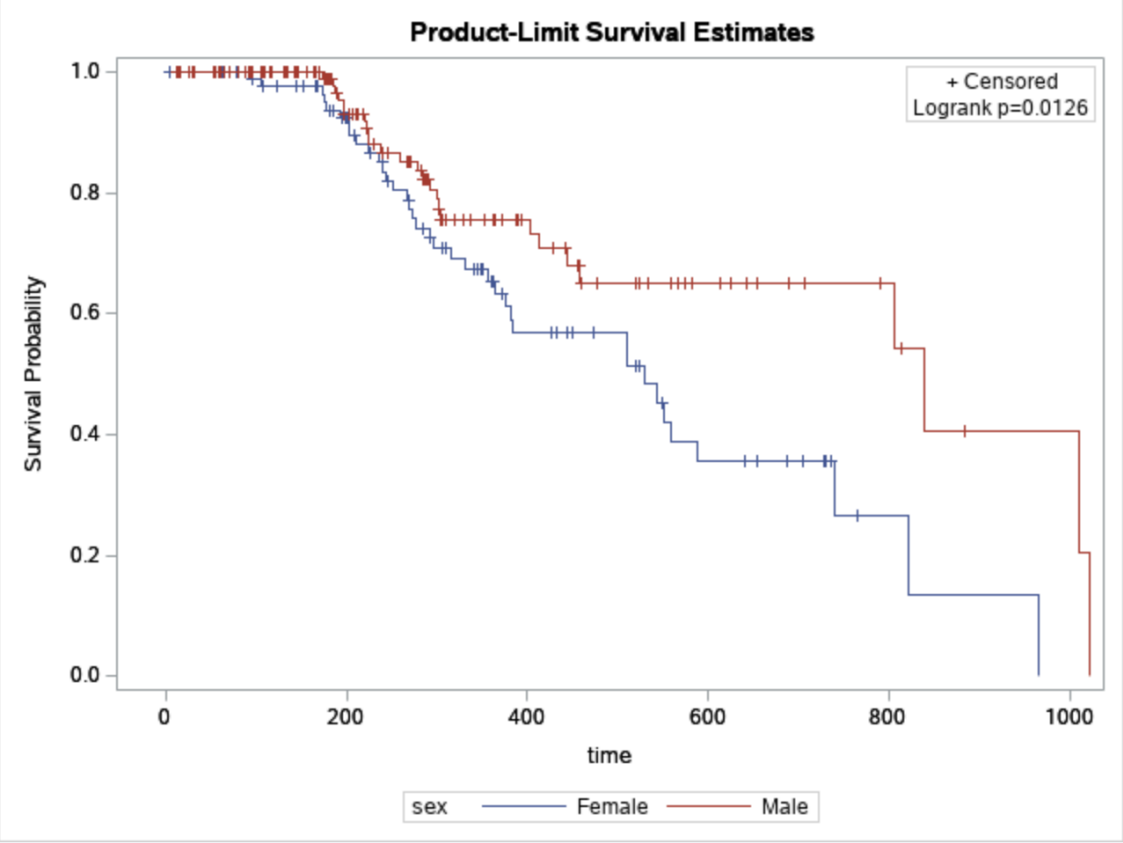

Figure 2: Comparison of survival experience between males and females

2.4 Semi-parametric Model (PH model)

2.4.1 Variable Selection and Stratification

In order to find the varaibles of central interest on the outcome, we applied three variable selection models here, backward elimination selection, stepwise selection and forward selection models. There are two ways of scoring a model based on its log-likelihood and complexity returned by PROC PHREG function in SAS here, Akaike Information Criterion (AIC) and Schwartz's Bayesian Criterion (SBC) (which is commonly known as the Bayesian Information Criterion (BIC)). As it is indicated in the figures below, the stepwise selection model comes with the AIC of 1009.889 with covariates (Table 6). Both the forward selection (Table 4) and backward selection model (Table 5) returned the AIC of 1009.462, which is pretty close to the value in stepwise. To develop the most appropriate model, we move on to compare the SBC indexes in three models. Both forward selection and backward selection model have SBC value of 1023.441 while the stepwise selection model have SBC value of 1018.276. Therefore, we determine to set up the results of utilize the stepwise selection as the final model.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1026.050	999.462
AIC	1026.050	1009.462
SBC	1026.050	1023.441

Table 4: Forward Selectioin Model Criterion

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1026.050	999.462
AIC	1026.050	1009.462
SBC	1026.050	1023.441

Table 5: Backward Selectioin Model Criterion

In the stepwise selection model, we set up the level of entry as 0.25 and the level of stay of 0.15 based on the common use. It turns out that the variables remaining in the final model are the sex, ECOG performance score, and Karnofsky performance score rated by physician (Table 7). The variable of weight loss in last six months was kicked out because its p-value exceeds 0.15 when it is included in the model.

As we have mentioned previously, the goal of this paper is to investigate the different impact on the outcome between two groups of sex, therefore we pre-specified to stratify by sex in the next model. The male group is represented by 1 and female by 2. Through PROC PHREG function in SAS, the results suggest that the risk of lung cancer in female is 0.588 times that in male (Table 8).

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	1026.050	1003.889
AIC	1026.050	1009.889
SBC	1026.050	1018.276

Table 6: Stepwise Selectioin Model Criterion

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	ph.ecog1		1	1	12.7198		0.0004
2	sex		1	2	6.7679		0.0093
3	ph.karno		1	3	2.5210		0.1123
4	wt.loss1		1	4	1.9138		0.1665
5		wt.loss1	1	3		1.9125	0.1667

Table 7: Stepwise Selectioin Model Results

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
sex	2	1	-0.53028	0.16718	10.0614	0.0015	0.588	sex 2

Table 8: Estimator Results Stratified by Sex

2.4.2 Model Checking

The Cox proportional hazards (PH) model makes two major assumptions. One of them is that the hazard functions for the survival curves of different strata will be proportional over the period of time t, and the other one is the relationship between the log of hazard $h(t)$ and each of the covariates will be linear. For this project, we omit the details for the latter assumption and focus on the first one as the proportional hazards assumption is significant in terms of the interpretations and the use of PH model. The following introduces some graphical methods, the interaction test, and residuals plot for assumption checking.

Graphical Approach One of the most popular strategies for PH assumption checking is to compare the survival curves visually. Recall for a PH model, $S(t|Z=z) = e^{-\int h_0(t)e^{\beta z}dt} = S_0(t)^{e^{\beta z}}$. After the log-log transformation (i.e., $log\{-logS(t|Z=z)\}$), we will have

$$log\{-log\hat{S}(t|Z=z)\} - log\{-log\hat{S}_0(t)\} = \beta,$$

where $Z \in \{0,1\}$ is an indicator variable, e.g., sex. Such equation implies the two curves, albeit relatively subjective, will be approximately paralleled if the proportional assumption holds. Figure 3 shows the transformed survival functions estimated by the K-M estimator along with the log of time (days). The other approach is to compare the differences between the observed KM estimates and fitted survival functions from the PH model as the Figure 4 shown. Both these two methods suppose `sex` is the only covariate in the model.

Residuals & Interaction Test Recommended by Hosmer et al[XX1]., the assessment of PH assumption includes two steps. Starting from evaluating the significance of each covariate through a global test (e.g., Score test, partial likelihood ratio test), and then confirm the results by checking the scatter plot of Schoenfeld residuals. For these methods, we add additional two continuous variables `meal.cal` and `wt.loss` and their corresponding interaction terms with time t to the original model. The reasons for choosing these two covariates is that the nutritional status and caloric intake are believed to be vital for the patients' survival as well as their response to the chemotherapy[1]. Results are shown in Figure 5 followed by part of the summary table from the interaction test.

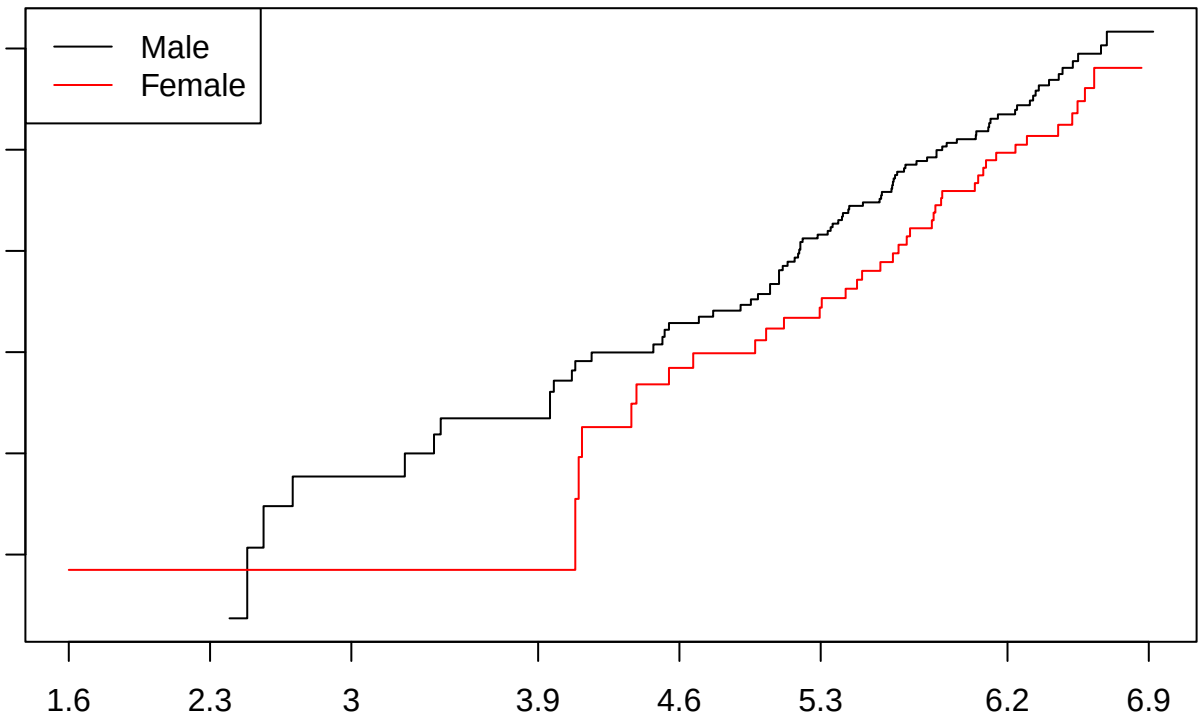

Figure 3: Log of Negative Log of Estimated Survival Functions

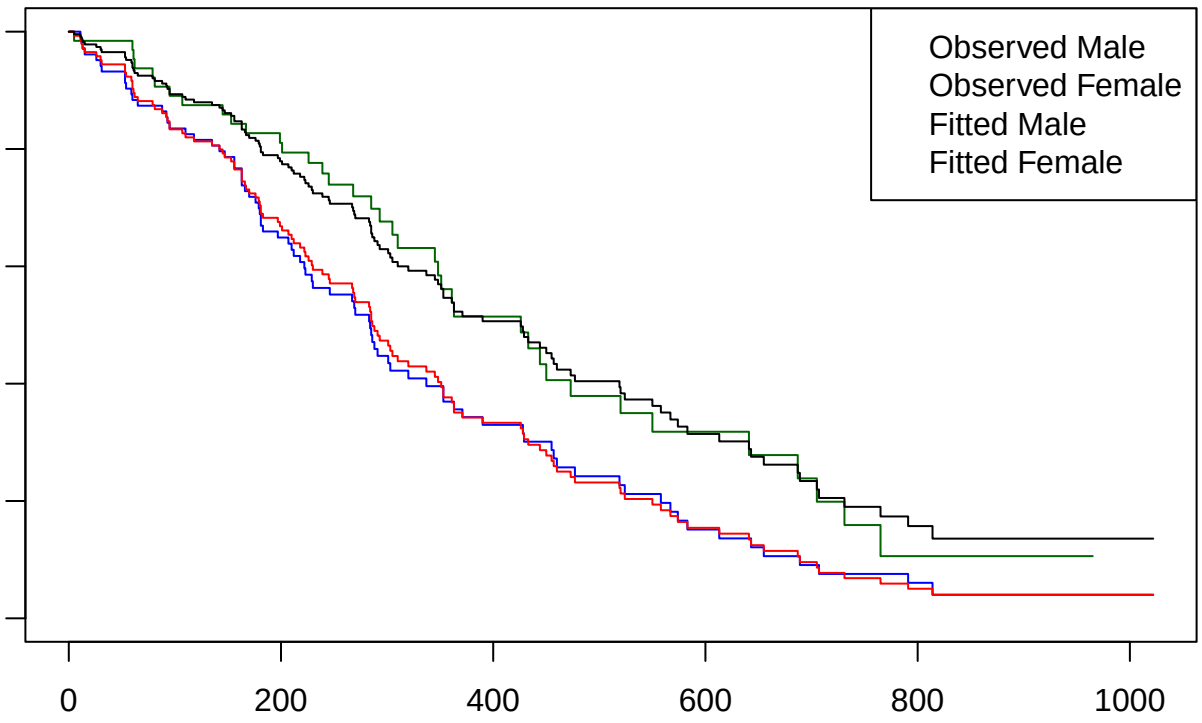

Figure 4: Observed vs. Fitted

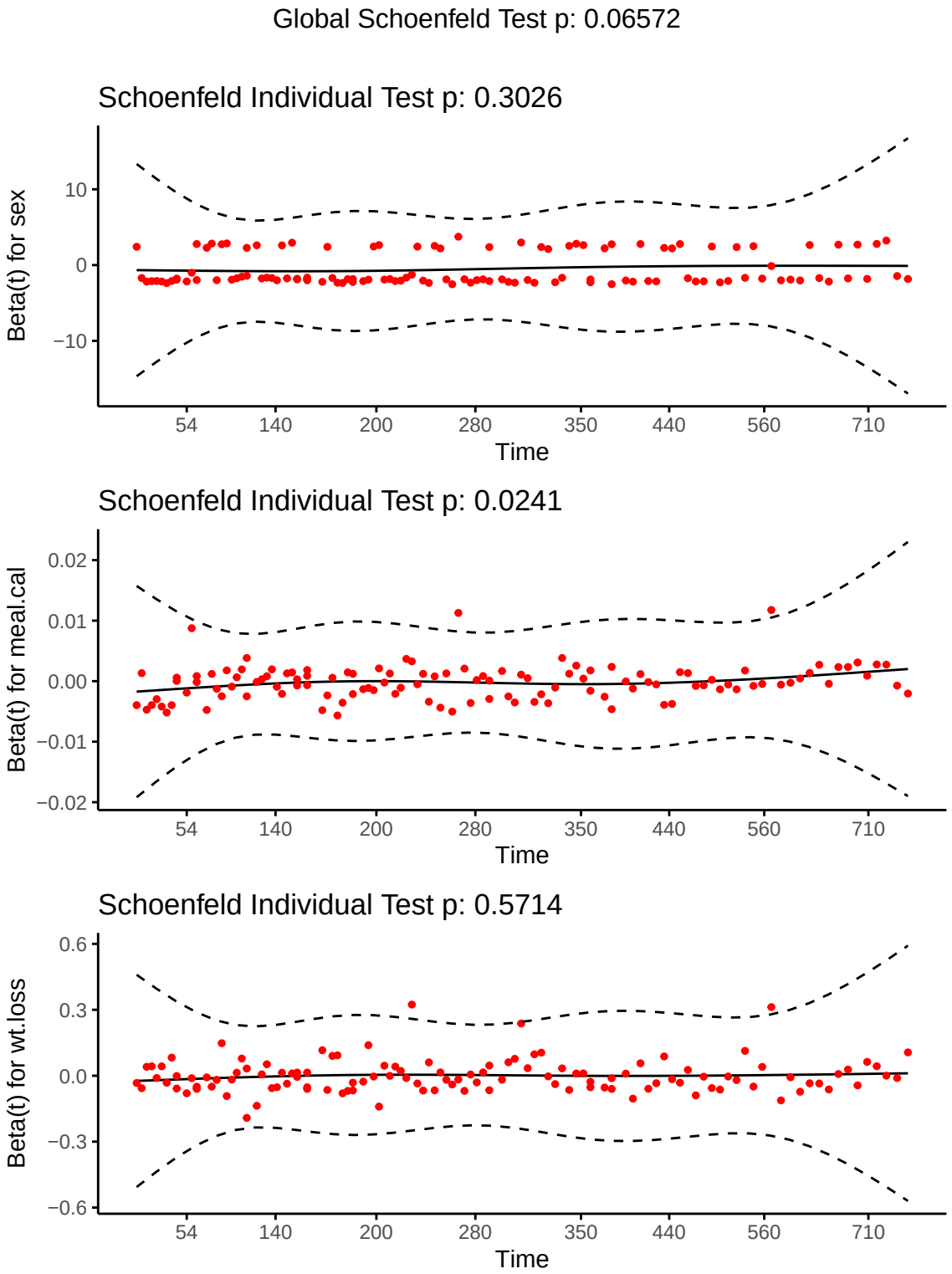

Figure 5: Schoenfeld Test

Table 9: Summary Table for Interaction Test (part)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
sex2	-0.405	0.667	0.479	-0.847	0.397
meal.cal	0.000	1.000	0.000	0.527	0.598
wt.loss	0.010	1.010	0.016	0.643	0.520
time	-1.408	0.245	0.245	-5.748	0.000
sex2:time	0.001	1.001	0.002	0.384	0.701
meal.cal:time	0.000	1.000	0.000	-0.439	0.661
wt.loss:time	0.000	1.000	0.000	-0.411	0.681

Figures from the graphical analysis demonstrate that the proportional hazards assumption is hold if there is only one indicator variable `sex` in the model. While the p values from the Schoenfeld test suggest that both `sex` and `wt.loss` fulfill the assumption, i.e., P>0.05. The violation of the assumption for `meal.cal` may be resolved by either adding a interaction term into the model or performing stratification.

2.5 Parametric Models

While the semi-parametric model focuses on the influence of covariates on hazard, a fully parametric model can calculate the distribution form of survival time. There are several advantages of parametric models. First, full maximum likelihood can be used to estimate parameters. Second, given a correct model assumption, the estimate can be more efficient and precise. Third, the model can be used to predict survival times, and residuals can represent the difference between observed and estimated values of time.

Based on the survival curves generated by the Kaplan-Meier estimator, we observed that the underlying probability distribution could possibly follow an exponential distribution or a Weibull distribution. Actually, both curves for females and males indicate some trend of increasing hazard rates, as they first concave down and then concave up, so we are more inclined to a Weibull distribution. Nonetheless, we first checked if the hazard functions both both sex are constant, as the correct choice of distribution is critical in fitting a parametric model.

2.5.1 Model Checking

Several plots can be used to check if the hazard rate is constant. First, we plotted $-log\hat{S}(t)$ against t to estimate the cumulative hazard function, and a straight line will indicate a constant hazard rate. Then, we referred to the previous plot of $log(-log\hat{S}(t))$ against $log\ t$ as shown in [figure x]. For an exponential distribution, $log(-logS(t)) = log\lambda t = log\lambda + logt$, and for a Weibull distribution, $log(-logS(t)) = log\lambda t^{\alpha} = log\lambda + \alpha logt$. Thus a straight line with a slope of 1 will indicate a constant hazard rate. At last, we fitted the survival curves into both an exponential and a Weibull distribution to visually inspect the similarity to the K-M estimators. The results are shown in [figure x]. s

[figure x]

For the $-log\hat{S}(t)$ plot, we can see that the curve for females is obviously non-linear, indicating a better choice of the Weibull distribution. For the $log(-log\hat{S}(t))$ plot, the slopes are larger than 1, especially for females, also preferring a Weibull distribution. In the last plot, it is also clear that a Weibull distribution fits the data more precisely than an exponential distribution. As a result, it is decided to use the Weibull distribution as the baseline hazard function to fit the parametric regression model.

2.5.2 Fitting Parametric Regression Models

For a Weibull baseline hazard function, the Accelerated Failure-Time Model is equivalent to the Proportional Hazard Model, so we only show the results for the fit of the PH Model, which gives better interpretations

of coefficients. After conducting a backward selection with significance level of 0.15, the model contains 4 significant covariates: `sex`, `ph.ecog`, `ph.karno`, `wt.loss`. The estimates of the coefficents, the hazard ratios and p-values are shown in [table x]. Similar to the results of the semi-parametric models: males have higher hazards compared to females; patients with higher ECOG performance scores rated by the physician have higher hazards; and although Karnofsky performance score rated by physician and weight loss in last six months are significant in p-values, the hazard ratios related to them is close to 1, thus their effects are actually not significant.

[table x]

2.5.3 Goodness of Fit

In order to check the goodness of fit, we compared the estimated baseline cumulative hazard function of the parametric model to that of a semi-parametric model, which is close to the observed data because no function is specified for the baseline hazard function. The results are shown in [figure x]. The solid curve is the parametric Weibull cumulative hazard function and the dashed curve is the Cox baseline cumulative hazard function. It appears that the parametric function fits well to the semi-parametric function.

[figure x]

3 Conclusion

For patients with advanced lung cancer, there is no significant difference in survival probability between males and females at early time points. However, the gap of survival probability widens over time. Proportional hazards assumption is hold if there is only one indicator variable sex in the PH model, and results show that females have 41% reduction of the risk of death than males. For the parametric model, Weibull distribution as the baseline hazard function is used. We can conclude that males have higher hazards compared to females, patients with higher ECOG performance scores rated by the physician have higher hazards, and the impact of Karnofsky performance score rated by physician and weight loss in last six months on the survival rate are not significant.

4 Discussion

In this project, we use non-parametric estimate, hypothesis testing, semi-parametric model and parametric model to conduct a survival analysis of patients with advanced lung cancer. However, there are some limitations in this project.

Missing Values We removed missing values in the dataset before statistical analysis. This is an easy way to deal with missing data but may lead to less accurate analysis results due to the reduced sample size. There are some other methods to keep the number of observations we have, such as mean imputation and regression imputation. But those methods may increase the complexity of the data manipulation process.

Linearity Between Log Hazard and Covariates In the assumptions checking of Cox proportional hazards (PH) model, we ignore the assumption that relationship between the log of hazard $h(t)$ and each of the covariates will be linear. However, in order to make sure the PH model is appropriate in this project, we need to further detect linearity between log hazard and the covariates, such as martingale residuals r_{Mi} or deviance residuals r_{Di} to assess the potential outliers.

Competing Risk

Multicovariates Analysis

Reference

[1] Loprinzi CL. Laurie JA. Wieand HS. Krook JE. Novotny PJ. Kugler JW. Bartel J. Law M. Bateman M. Klatt NE. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*. 12(3):601-7, 1994.

[2] Preston SH, Heuveline P, Guillot M. Demography: measuring and modeling population processes. Blackwell Publishers, 2001.

[3] Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. International journal of Ayurveda research, 1(4), 274.

[XX1] Hosmer, D. W., Lemeshow, S., & May, S. (2011). In Applied survival analysis: Regression modeling of time to event data. essay, Wiley.

Appendix

For codes please click *here*.