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
Asymptomatic	63 (28%)	26 (41%)	37 (23%)	
Symptomatic but completely ambulatory	113 (50%)	31 (49%)	82 (50%)	
In bed <50% of the day	50 (22%)	6 (9.5%)	44 (27%)	
In bed > 50% of the day but not bedbound	1 (0.4%)	0 (0%)	1 (0.6%)	
Bedbound	0 (0%)	0 (0%)	0 (0%)	
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 1 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

Figure 1: Table 1-1 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

Figure 2: Table 1-2 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.

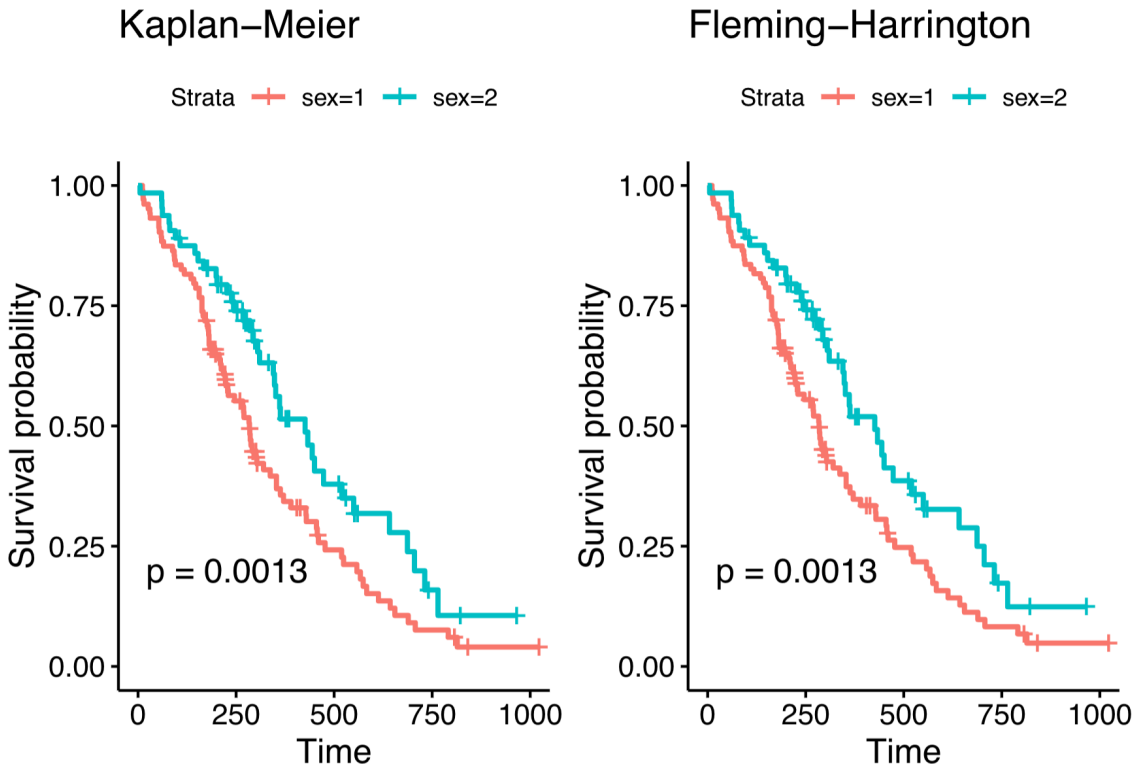


Figure 3: Kaplan-Meier model and Fleming -Harrington model (sex=1 (male), sex=2 (female))

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. (Figure 2).

2.3 Hypothesis Testing

2.4 Semi-parametric Model (PH model)

2.4.1 Variable Selection and Stratification

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 will be paralleled if the proportional assumption holds. Figure 4 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 5 shown. Both these two methods suppose `sex` is the only covariate in the model.

Residuals & Interaction Test XXX

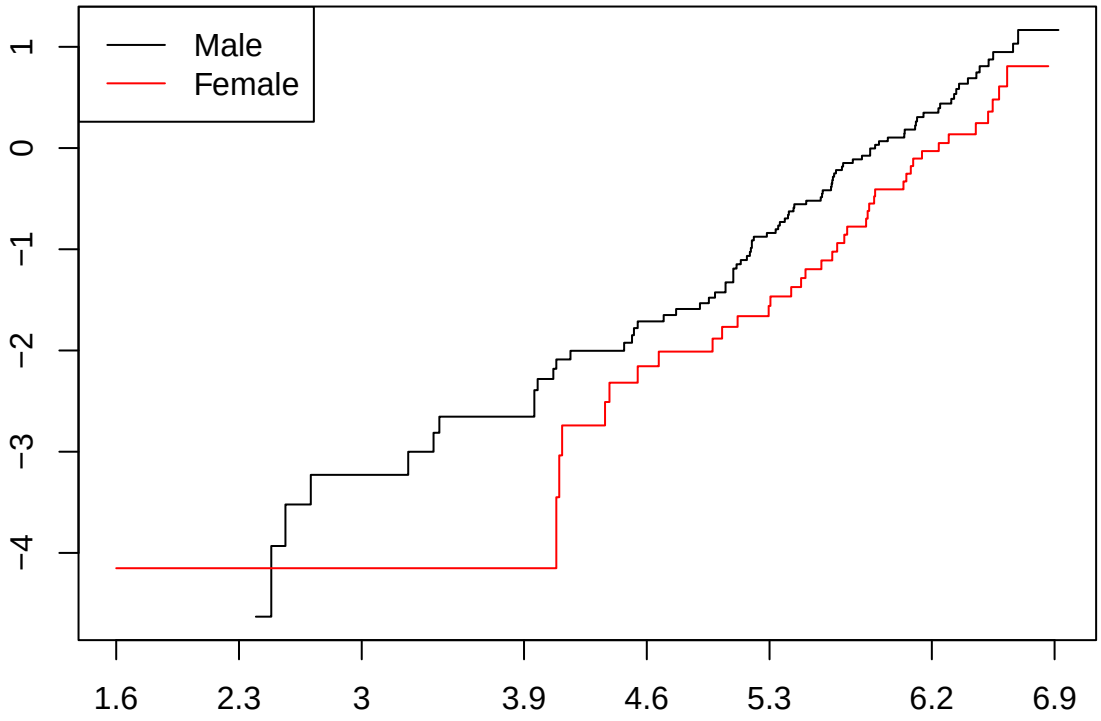


Figure 4: Log of Negative Log of Estimated Survival Functions

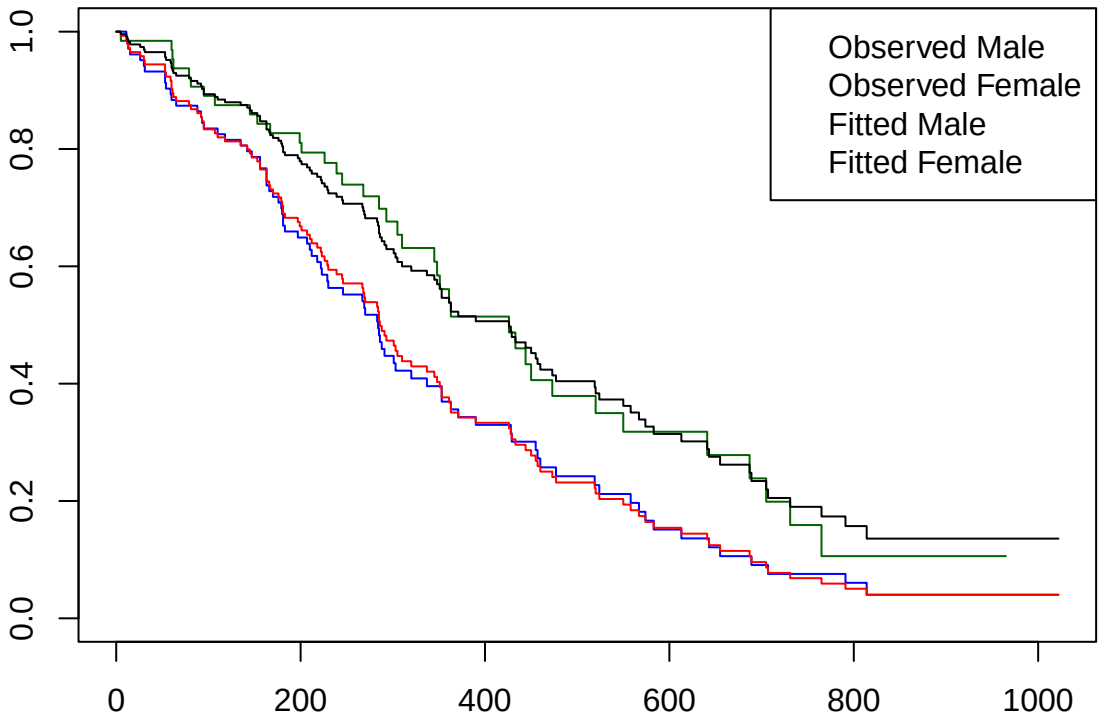
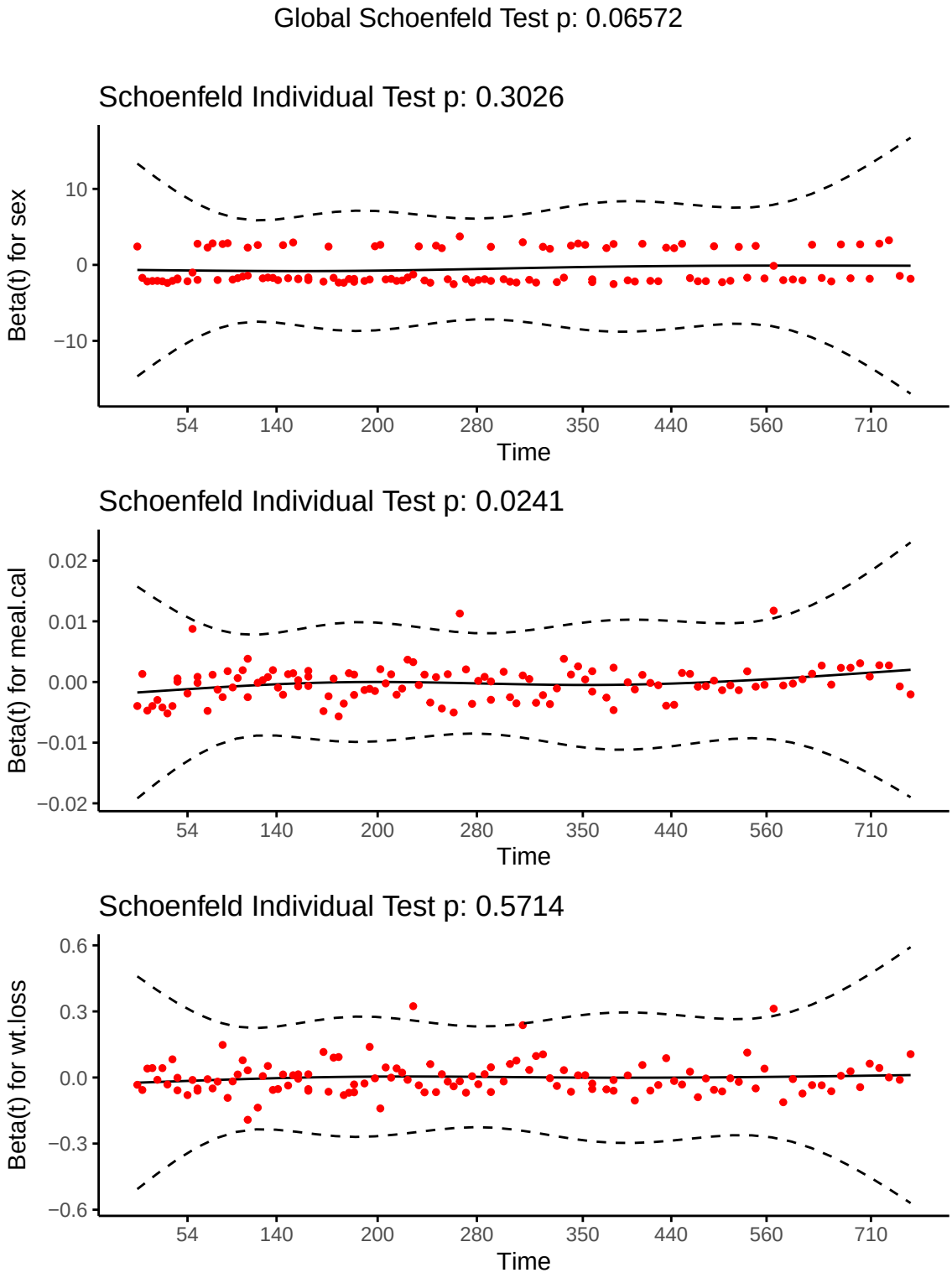


Figure 5: Observed vs. Fitted




```
## Call:
## coxph(formula = Surv(time, status == 2) ~ sex + meal.cal + wt.loss +
##     sex * time + meal.cal * time + wt.loss * time, data = dat_lung)
##
##   n= 167, number of events= 120
##
```

```
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## sex2          -4.053e-01  6.668e-01  4.787e-01 -0.847    0.397
## meal.cal       2.345e-04  1.000e+00  4.447e-04  0.527    0.598
## wt.loss        1.041e-02  1.010e+00  1.619e-02  0.643    0.520
## time          -1.408e+00  2.446e-01  2.450e-01 -5.748 9.02e-09 ***
## sex2:time      6.079e-04  1.001e+00  1.583e-03  0.384    0.701
## meal.cal:time -7.809e-07  1.000e+00  1.778e-06 -0.439    0.661
## wt.loss:time  -2.643e-05  1.000e+00  6.428e-05 -0.411    0.681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## sex2             0.6668     1.4998    0.2609    1.7039
## meal.cal         1.0002     0.9998    0.9994    1.0011
## wt.loss          1.0105     0.9896    0.9789    1.0430
## time             0.2446     4.0885    0.1513    0.3953
## sex2:time        1.0006     0.9994    0.9975    1.0037
## meal.cal:time    1.0000     1.0000    1.0000    1.0000
## wt.loss:time     1.0000     1.0000    0.9998    1.0001
##
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 973.8  on 7 df,   p=<2e-16
## Wald test            = 34.96  on 7 df,   p=1e-05
## Score (logrank) test = 184.4  on 7 df,   p=<2e-16
```

The above figures demonstrate that the proportional hazards assumption is hold given there is only one indicator variable `sex` in the model. From the above figures, the p values for both variables `sex` and `wt.loss` are greater than 0.05 except the variable `meal.cal`, which means that the proportional hazard assumptions is violated only for `meal.cal`. This result is not consistent with the interaction test where the assumption is retained for all three variables (P>0.05). XXX recommended a two-step procedure for PH assumption assessment where calculating the

2.5 Parametric Models

3 Conclusion

4 Discussion

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.

Appendix

For codes please click *here*.