Survival Analysis Group 5 Project

Chloe Jian, Hening Cui, Jibei Zheng,
Pengchen Wang, Xueqing Huang, Qihang Wu

December 05, 2022

Outline

Introduction

Motivation: Lung cancer is a disease with a very high prevalence. Prognostic factors provides important information for patients with cancer. A better understanding of patients' prognosis can help make appropriate therapeutic decisions.

Data Source: Lung cancer dataset in Survival package of R

Purpose: Survival estimation, stratified by sex

- 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 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

Exploratory Data Analysis

Table: Patient Characteristics

Variable	**Overall**, N = 228[1]	**Alive**, N = 63[1]	**Death**, N = 165[1]	**p-value**[2]
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%)	

[1] Mean (SD); n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

Life Table

Lifetable for Male

	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

Life Table

Lifetable for female

	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

50% for male: 285-286 days

50% for female: 433-434 days

Kaplan-Meier and Fleming-Harrington model

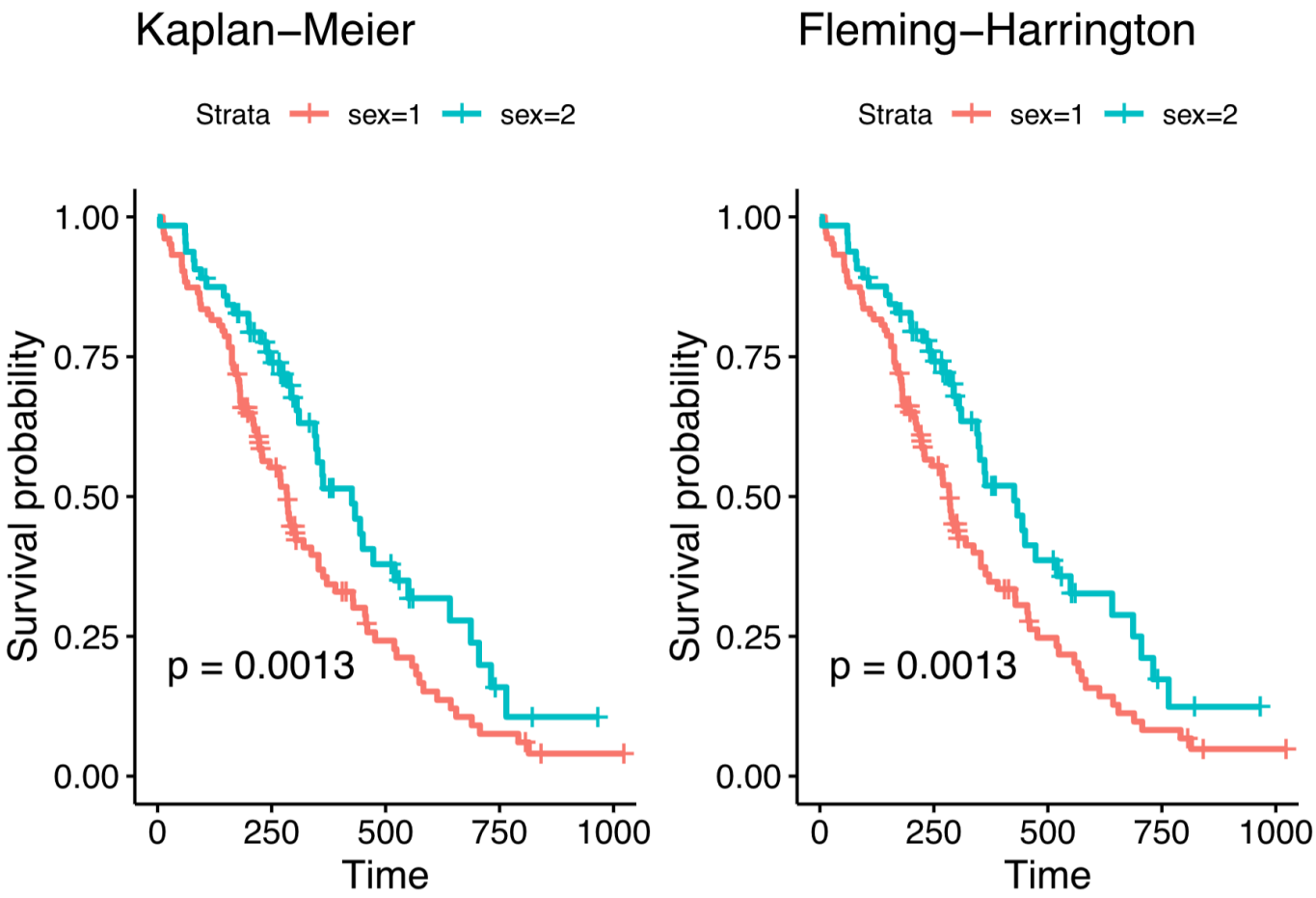


Figure: Kaplan-Meier model and Fleming -Harrington model (sex=1(male), sex=2(female))

Kaplan-Meier and Fleming-Harrington model

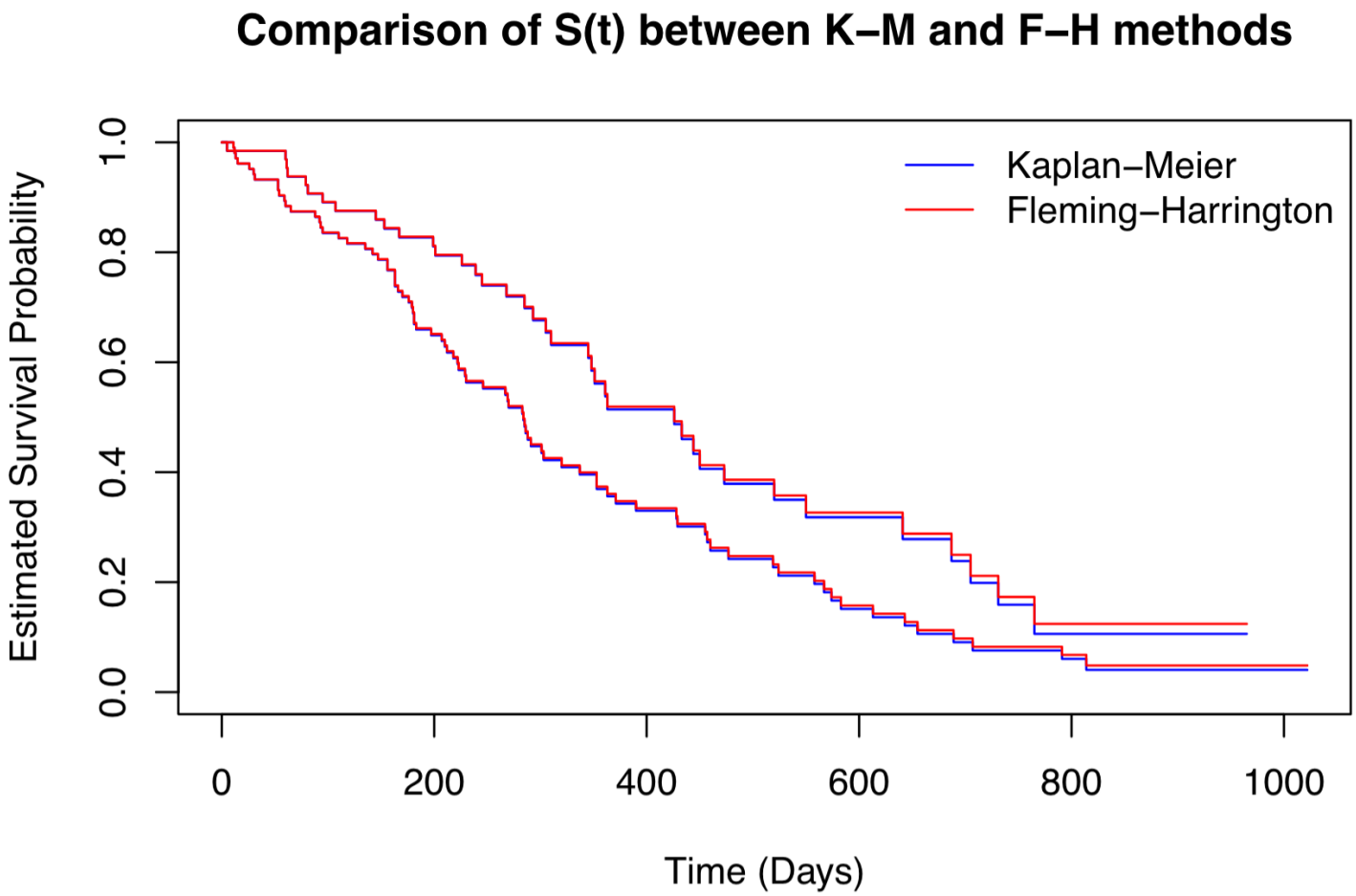


Figure: Kaplan-Meier model vs Fleming-Harrington model

Non-parametric test

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

Figure: Compare survival experience between males and females

Survival Curve

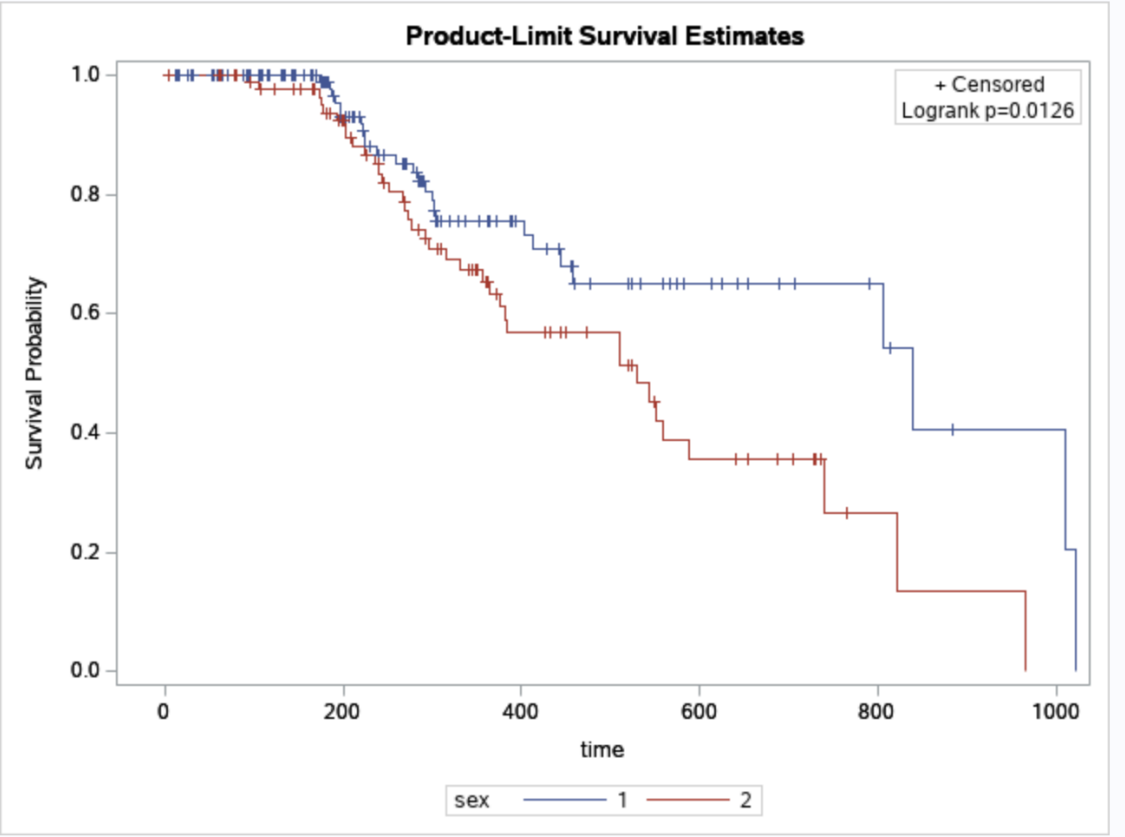


Figure: Compare survival experience between males and females

Variable Selection - Stepwise Selection

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
sex	2	1	-0.52581	0.19684	7.1354	0.0076	0.591	sex 2
ph.ecog1		1	0.74026	0.21055	12.3607	0.0004	2.096	
ph.karno		1	0.01766	0.01113	2.5188	0.1125	1.018	

Analysis of Effects Eligible for Removal			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	7.1354	0.0076
ph.ecog1	1	12.3607	0.0004
ph.karno	1	2.5188	0.1125

Analysis of Effects Eligible for Entry			
Effect	DF	Score Chi-Square	Pr > ChiSq
age	1	1.2229	0.2688
pat.karno	1	1.4557	0.2276
meal.cal1	1	0.1002	0.7516
wt.loss1	1	1.9138	0.1665

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
5.1428	4	0.2730

Note: Model building terminates because the effect to be entered is the effect that was removed in the last step.

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

Step 5. Effect wt.loss1 is removed. The model contains the following effects:

sex ph.ecog1 ph.karno

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

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

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	22.1611	3	<.0001
Score	22.2893	3	<.0001
Wald	21.9435	3	<.0001

Variable Selection - Forward and Backward Selection

Forward Selection

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

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
sex	2	1	-0.54760	0.19832	7.6246	0.0058	0.578	sex 2
ph.ecog1		1	0.72880	0.22576	10.4214	0.0012	2.073	
ph.karno		1	0.02029	0.01110	3.3432	0.0675	1.021	
pat.karno		1	-0.01236	0.00793	2.4310	0.1190	0.988	
wt.loss1		1	-0.01323	0.00794	2.7815	0.0954	0.987	

Backward Selection

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

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
sex	2	1	-0.54760	0.19832	7.6246	0.0058	0.578	sex 2
ph.ecog1		1	0.72880	0.22576	10.4214	0.0012	2.073	
ph.karno		1	0.02029	0.01110	3.3432	0.0675	1.021	
pat.karno		1	-0.01236	0.00793	2.4310	0.1190	0.988	
wt.loss1		1	-0.01323	0.00794	2.7815	0.0954	0.987	

Variable Stratification

- Our goal is to investigate the difference between two groups of sex, therefore we pre-specified to stratify by sex in the model.
- sex: Male=1, Female=2

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

- We do find that the risk of lung cancer in female is 0.588 times that in male.

Graphical Methods

Recall a PH model, $S(t|Z=z) = e^{-\int h_0(t)e^{\beta z}dt} = S_0(t)^{e^{\beta z}}$, by using a log-log transformation, i.e., $log\{-logS(t|Z=z)\}$, we have

$$log\{-log\hat{S}(t|Z=1)\} - log\{-log\hat{S}_0(t)\} = \beta \text{ for indicator variable Z.}$$

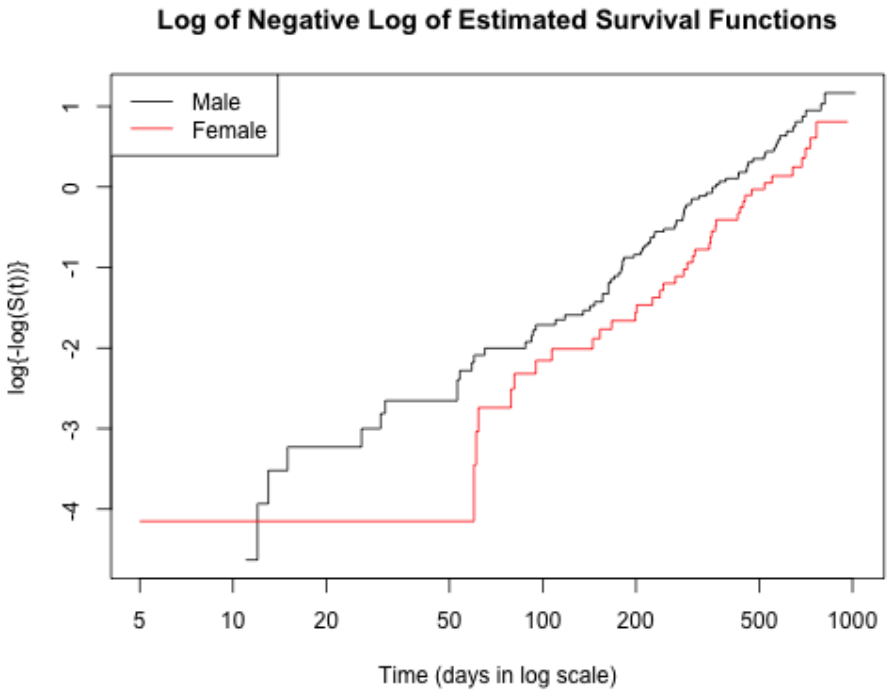
[1] Figure on the right represents the differences of KM estimate and the fitted PH model.

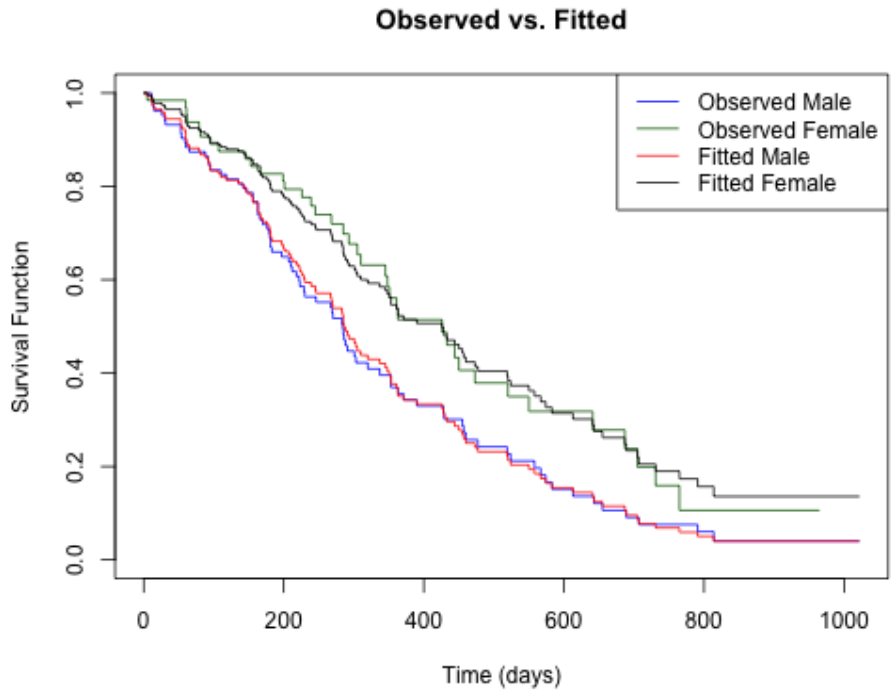
Graphical Methods

Recall a PH model, $S(t|Z=z) = e^{-\int h_0(t)e^{\beta z}dt} = S_0(t)^{e^{\beta z}}$, by using a log-log transformation, i.e., $log\{-logS(t|Z=z)\}$, we have

$$log\{-log\hat{S}(t|Z=1)\} - log\{-log\hat{S}_0(t)\} = \beta \text{ for indicator variable Z.}$$

This indicates two **parallel** lines under proportionality assumption. [1]



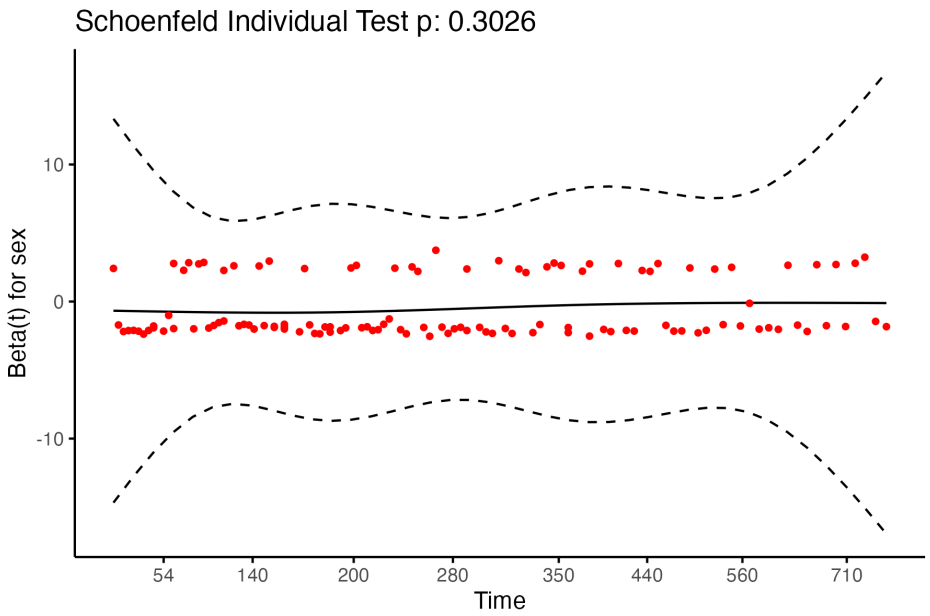


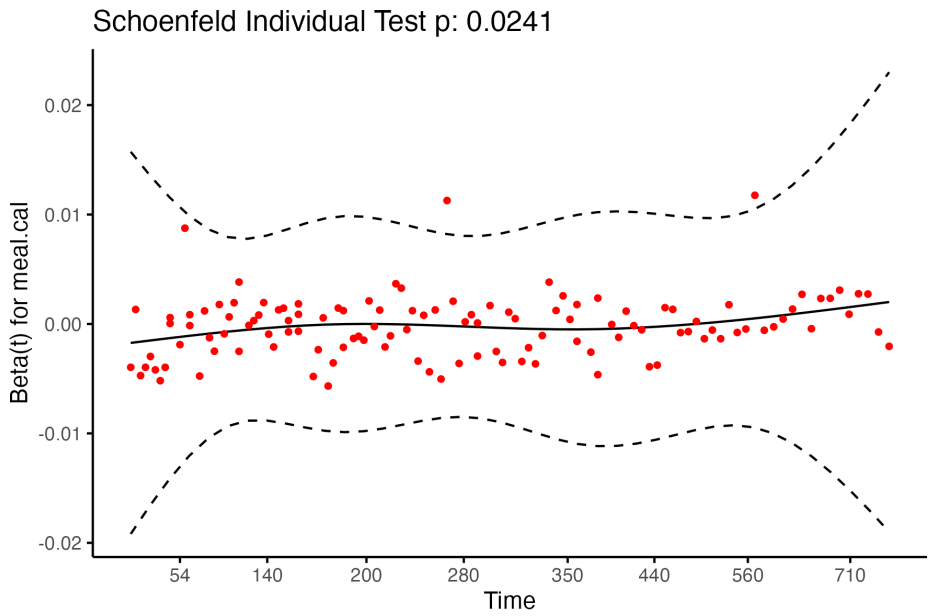
[1] Figure on the right represents the differences of KM estimate and the fitted PH model.

Schoenfeld Residuals

- The above methods work only for categorical variable, the slope in plots of residuals such as **Schoenfeld vs. time** can be used instead for continuous cases.
- Figures below based on a PH model, i.e., [2]

$$h(t|Z = \mathbf{z}) = h_0(t)e^{\beta_1 sex + \beta_2 meal.cal + \beta_3 wt.loss}$$





[2] Scaled Schoenfeld residuals for wt.loss are omitted here as the p value is also greater than 0.05.

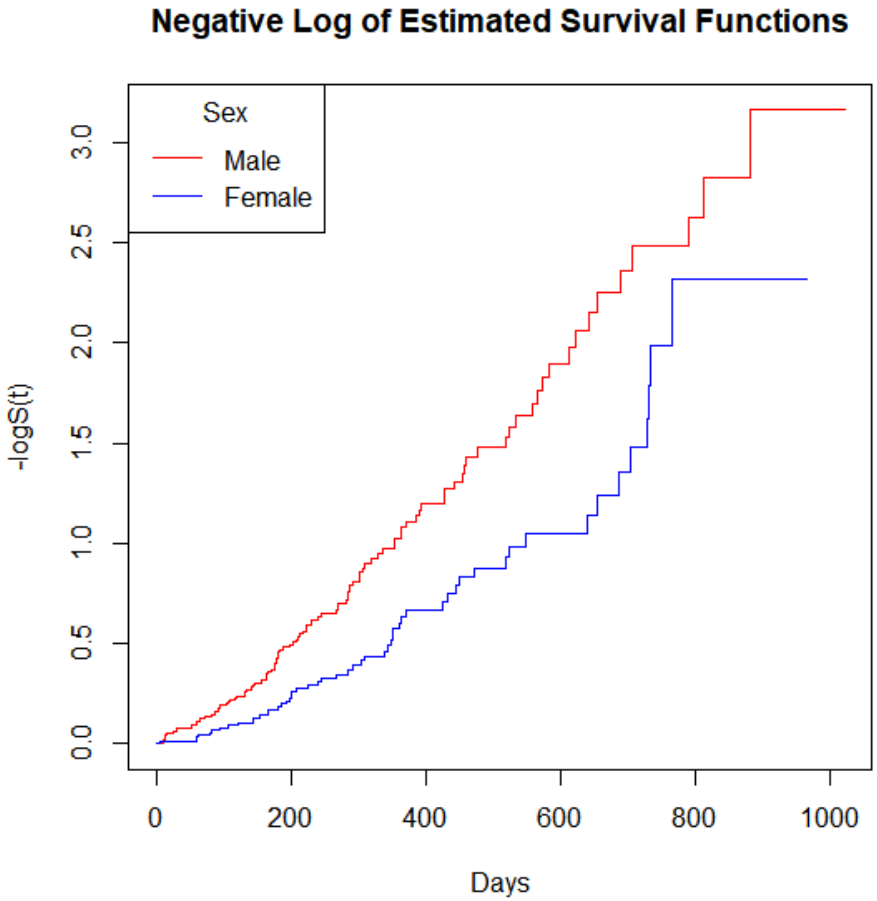
Interaction Test

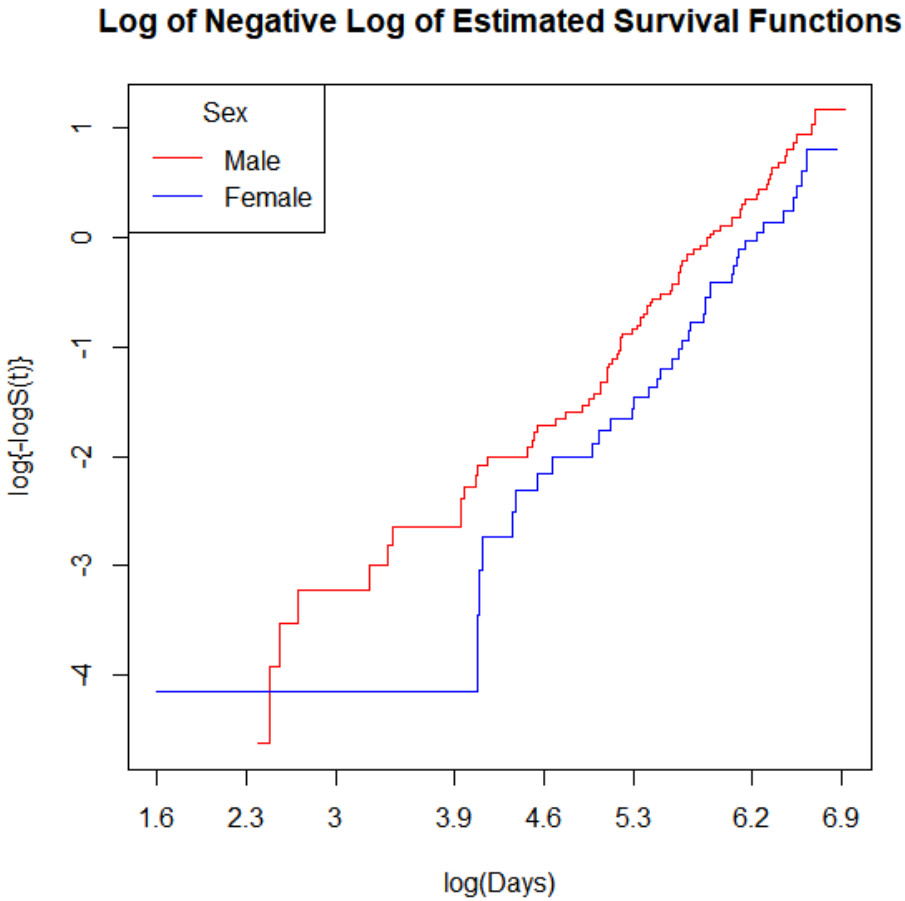
Using the same covariates as above, plus the corresponding interaction terms s.t., **covariate * time**. Results are shown as follows.

```
## 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
##
```

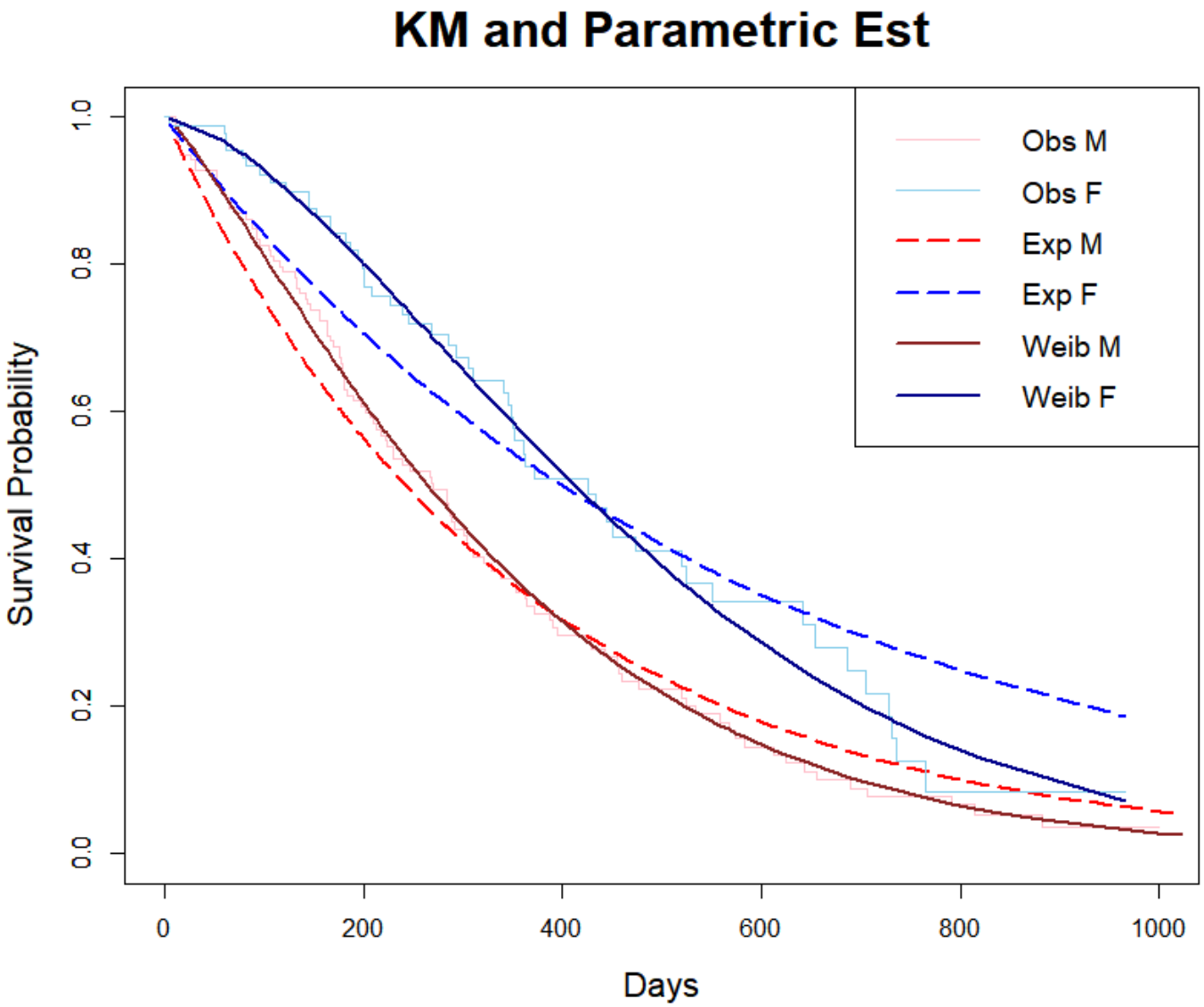
Figure: Interaction Test Summary (part)

Parametric Model Checking





Parametric Model Checking



Parametric PH Model

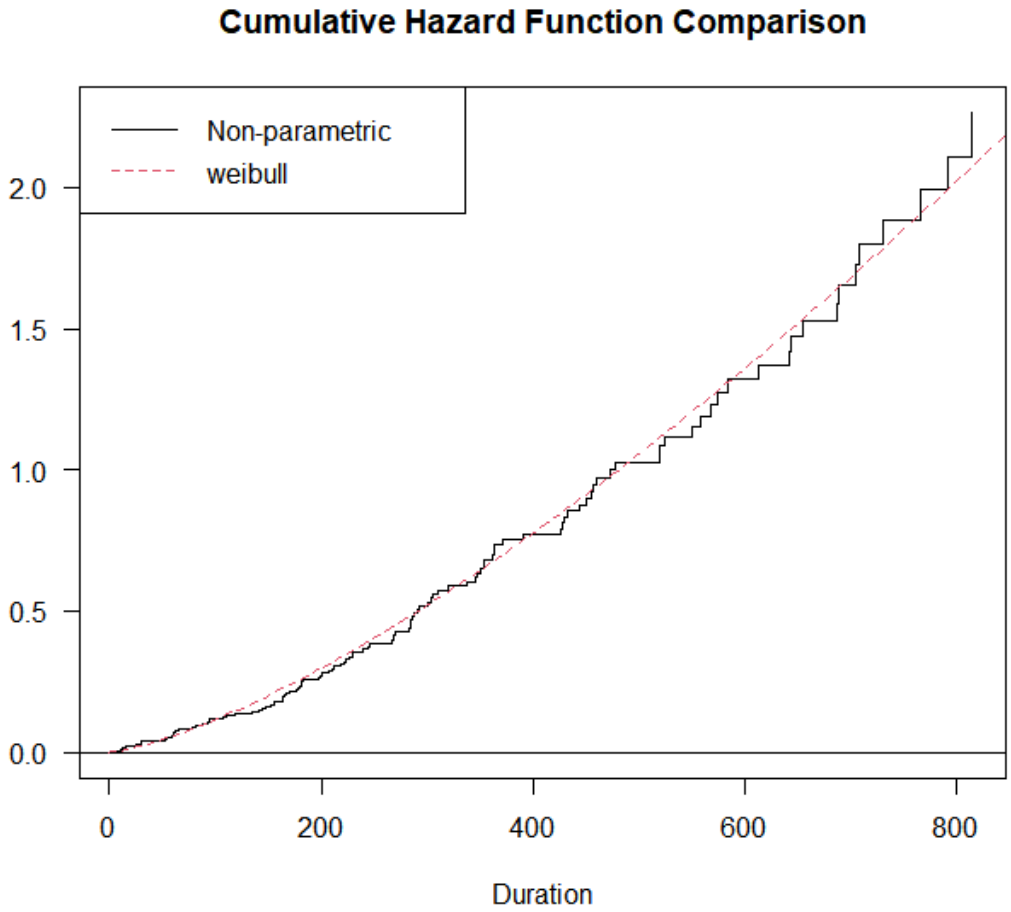
- Using Weibull distribution
- Significant variables ($\alpha = 0.15$): sex, ph.ecog, ph.karno, wt.loss

Covariate		Mean	Coef	Rel.Risk	S.E.	LR p
sex						0.0032
	1	0.579	0	1 (reference)		
	2	0.421	-0.571	0.565	0.199	
ph.ecog						0.0005
	0	0.309	0	1 (reference)		
	1	0.520	0.642	1.900	0.279	
	2	0.169	1.720	5.586	0.436	
	3	0.002	2.884	17.891	1.115	
ph.karno		82.850	0.020	1.020	0.011	0.0568
wt.loss		10.019	-0.012	0.988	0.008	0.1025

Events	120
Total time at risk	51759
Max. log. likelihood	-827.68
LR test statistic	26.78
Degrees of freedom	6
Overall p-value	0.000159438

Goodness of Fit

Comparison between parametric and semi-parametric PH regression models using the same variables.



Conclusion

Non-parametric Estimate(Lifetable/KM/FH):
Males have shorter survival time than females.

Hypothesis Testing(Log-rank/Likelihood Ratio Test):
Survival time is significantly different between males and females.

Semi-parametric Model (PH model):

- Stepwise selection: ph.ecog, sex and ph.karno.
- Model checking: Proportionality assumptions hold for both **sex**, **wt.loss**. The Schoenfeld residuals and interaction test render different conclusions for the covariate **meal.cal**.

Parametric Model:
Using Weibull distribution with significant variables sex and ph.ecog.

Discussion

- Processing for missing values: remove, multiple imputation, etc.
- Patients' performance scores are highly subjective.
- Detecting linearity between log hazard and the covariates, e.g., Martingale residuals r_{Mi} or deviance residuals r_{Di} to assess the potential outliers, etc.
- Source for competing risk analysis.
- Analysis for other covariates, e.g., ph.ecog.
- ...

Thank you for listening!

Reference

- 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.
- Preston SH, Heuveline P, Guillot M. Demography: measuring and modeling population processes. *Blackwell Publishers*, 2001.
- Goel, M. K., Khanna, P., Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, 1(4), 274.