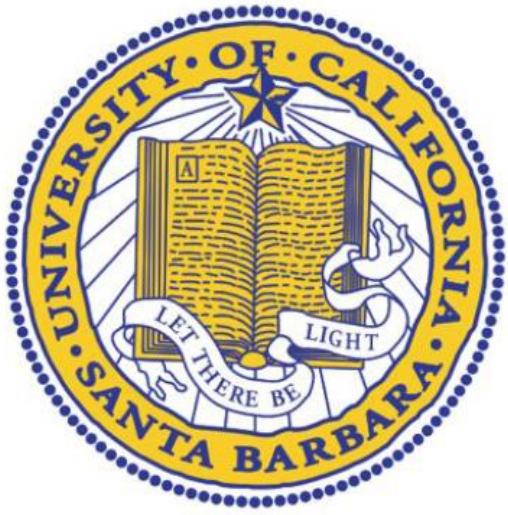


University of California, Santa Barbara
Department of Probability and Statistics
PSTAT 175 Final Project

Survival Analysis: Advanced Lung Cancer

Report by: Yeqi Zhang, Mujie Wang, Xuefeng Wu,
Zhisen Cai

Instructor: Professor Tashman

Data: 2019/12/6

Abstract

In the modern way of life, lung cancer is one of the most deadly malignancies. Variables such as karnofsky performance score and ECOG performance score may affect lung cancer patients' survival time. In this project, we will build a Cox Proportional Hazards model to study the relationship between these variables and survival model.

Data Source and Background Information

The dataset we used is NCCTG Lung Cancer Data, which describes survival in patients with advanced lung cancer from the North Central Cancer Treatment Groupis(Loprinzi et al. 1994). The data can be obtained from the R documentation: https://www.rdocumentation.org/packages/survival/versions/3.1-7/topics/lung. The variables in the data are as follows:

- Time (Survival time in days).
- Status (censoring status), coded as 1=censored, 2=dead.
- Age (Age in years).
- sex, coded as 1=Male, 2=Female.
- ph.ecog (ECOG performance score). It describes a patient's level of functioning in term of their ability to care for themselves, containing 6 levels, where "0" represents a fully active and healthy person and extendedly "5" represents a dead person

- ph.karno (Karnofsky performance score), coded from 0=bad to 100=good rated by physician, containing 11 levels. Karnofsky performance score is a standard way to describe the ability of patient to perform ordinary task. Karnofsky performance status scores range from 0 to 100 (0,10,20,etc). A higher score means the patient is better able to carry out daily activities. Ph.karno is the karnofsky performance score rated by physicans
- pat.karno (Karnofsky performance score as rated by patient). It functions as ph.karno but the score is decided by patients themselves.
- meal.cal (Calories consumed at meals).
- wt.loss it is the patient's weight loss in last six months

Research Question

We are interested in whether age, sex, ph.ecog. ph.karno, pat.karno, meal.cal and wt.loss would lead to death and how do they affect they affect the demise. In addition, we want to see whether there is any interaction between any of these covariates mentioned above.

Moreover, since we use the coxph as our basic model, we want to have an alternative way, AFT, to explore the data.

Data Preprocessing and Exploration

Before we build the Cox PH model, we try to understand the dataset first. By using dim and summary function we can get information that the original dataset has 228 observations and variable meal.cal has more than 40 missing values. Hence we want to discard meal.cal in our dataset since it loses 47/228=20.6% data.

```
      inst             time           status           age       
 Min.   : 1.0   Min.   :   5   Min.   :1.00   Min.   :39.0  
 1st Qu.: 3.0   1st Qu.: 167   1st Qu.:1.00   1st Qu.:56.0  
 Median :11.0   Median : 256   Median :2.00   Median :63.0  
 Mean   :11.1   Mean   : 305   Mean   :1.72   Mean   :62.4  
 3rd Qu.:16.0   3rd Qu.: 396   3rd Qu.:2.00   3rd Qu.:69.0  
 Max.   :33.0   Max.   :1022   Max.   :2.00   Max.   :82.0  
 NA's   :1                                                  
      sex           ph.ecog         ph.karno        pat.karno     
 Min.   :1.00   Min.   :0.000   Min.   : 50.0   Min.   : 30  
 1st Qu.:1.00   1st Qu.:0.000   1st Qu.: 75.0   1st Qu.: 70  
 Median :1.00   Median :1.000   Median : 80.0   Median : 80  
 Mean   :1.39   Mean   :0.952   Mean   : 81.9   Mean   : 80  
 3rd Qu.:2.00   3rd Qu.:1.000   3rd Qu.: 90.0   3rd Qu.: 90  
 Max.   :2.00   Max.   :3.000   Max.   :100.0   Max.   :100  
                NA's   :1       NA's   :1       NA's   :3    
    meal.cal        wt.loss       
 Min.   :  96   Min.   :-24.00  
 1st Qu.: 635   1st Qu.:  0.00  
 Median : 975   Median :  7.00  
 Mean   : 929   Mean   :  9.83  
 3rd Qu.:1150   3rd Qu.: 15.75  
 Max.   :2600   Max.   : 68.00  
 NA's   :47     NA's   :14      
```

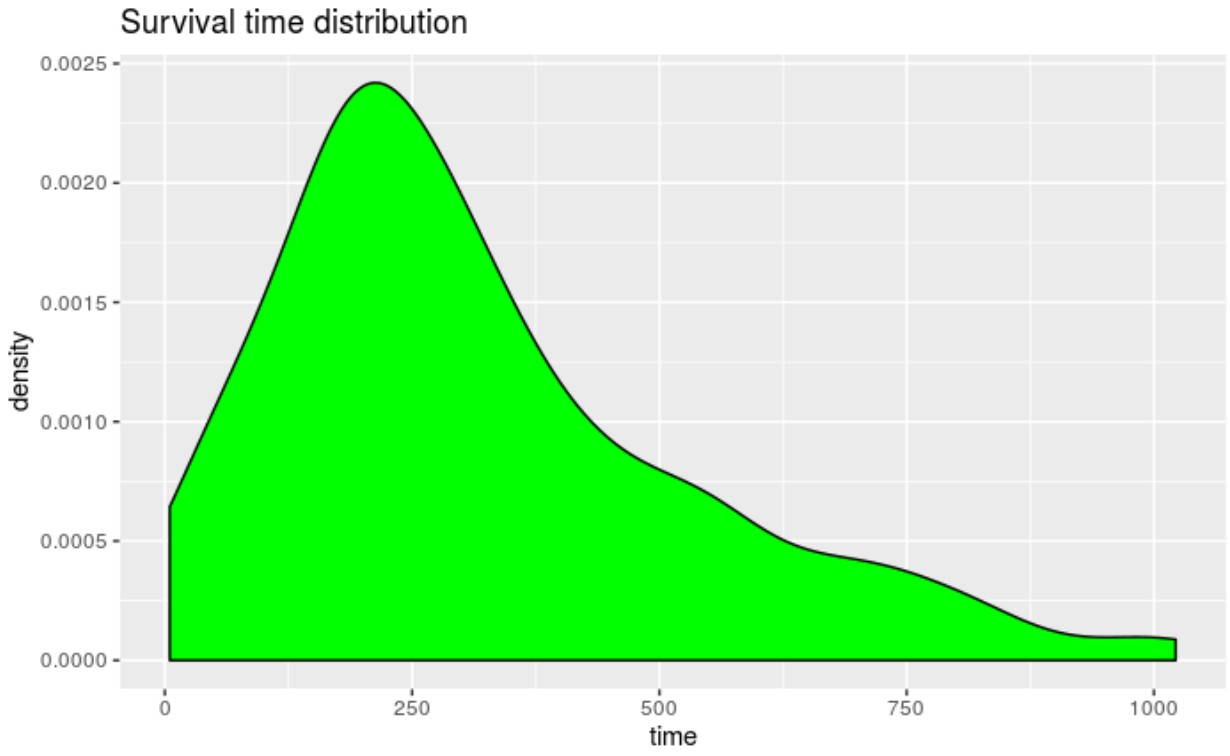
We then drop observations with other missing values in the dataset. And find out that it contains 209 complete observations. We also want to regroup ph.karno, pat.karno, age and wt.loss to help enhance the significance of our study since they all have a large range.

- For ph.karno and pat..karno, we find that if the patient's karnofsky performance is equal to or lower than 60, he must be taken care of by other people. What's more, they are unsuitable to have a chemotherapy or major surgery. Patient with karnofsky performance score greater and equal to 70 can live as a normal person.

Therefore, we decide to split into 2 groups of patients, patients with karnofsky performance score lower or equal to 60 and higher than or equal to 70.

- For wt.loss, since an unexplained weight loss equal or more than 10 should be considered a problem. Hence we separate weight loss to 2 groups, smaller than 10 and greater and equal to 10.
- For age we want to separate into 2 group, patient's age smaller than 60 and, that greater and equal to 60

Then, here is the overview of our final data



By applying the function quantile in R, we can see that 62.5% of the observations survive less than 320 days.

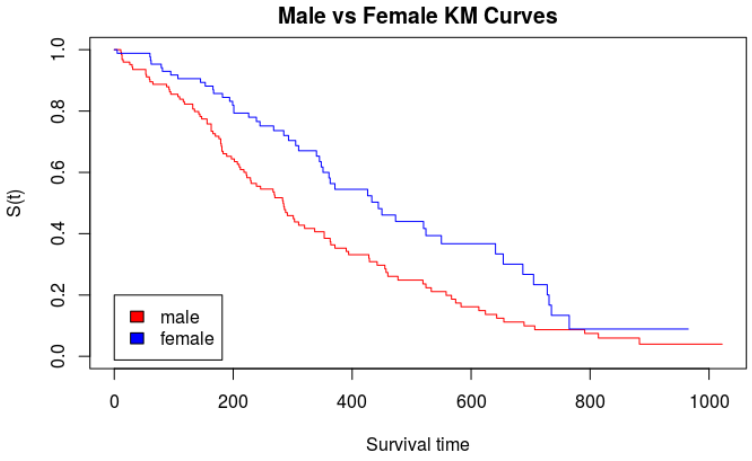
```
 0% 12.5%   25% 37.5%   50% 62.5%   75% 87.5%  100%
  5   105   176   211   269   320   429   574  1022
```

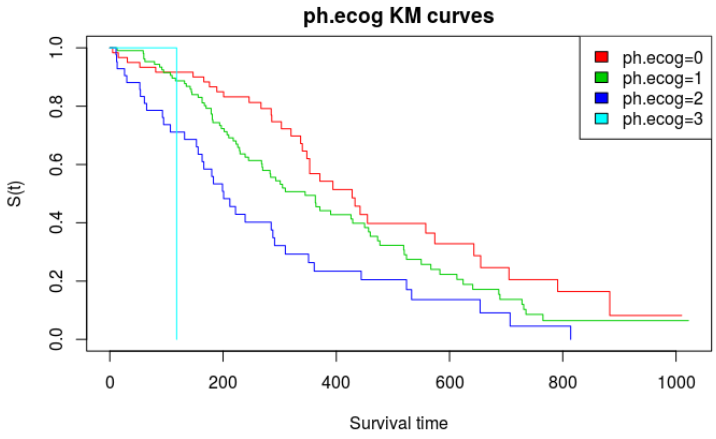
Model Building

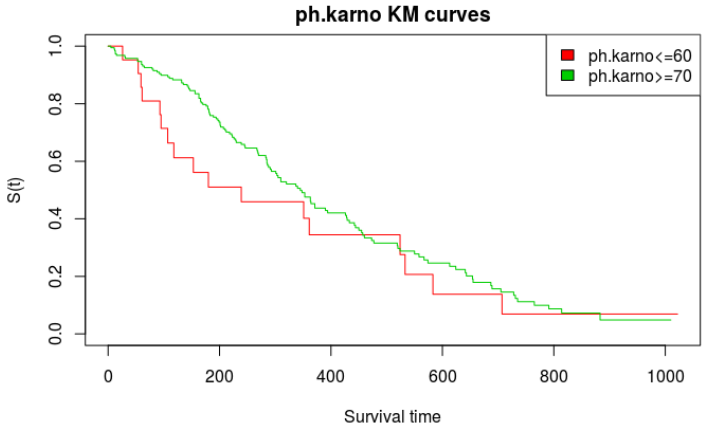
Now, we start to build our Cox PH model using the following variables: age, sex, ph.ecog, ph.karno, pat.karno and wt.loss,

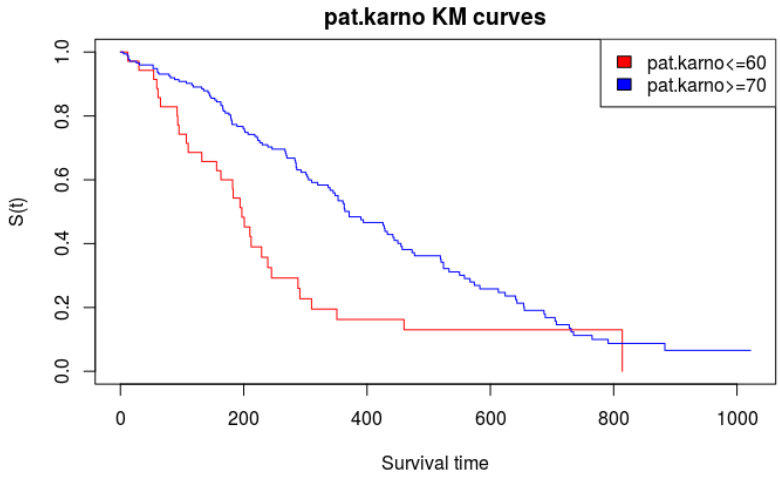
◆ Kaplan-Meier estimation curves

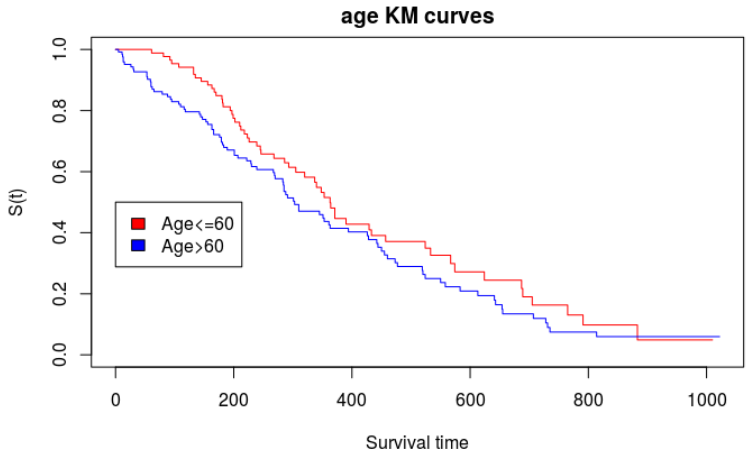
We plot the Kaplan-Meier survival curves to visually analyze the effects of each covariate.

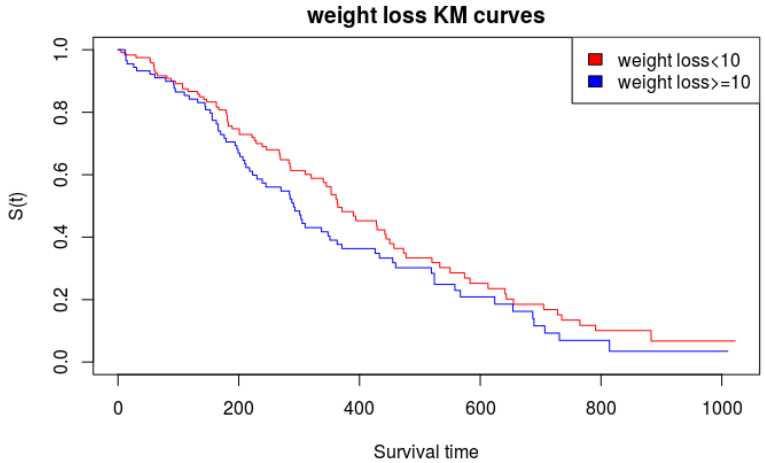












Looking at the plot above, we can conclude that sex, ph.ecog and pat.karno have influence on the survival probabilities since they have great difference in their graph while ph.karno, age and weight loss seem to be the same. We find that male Patient

seem to die earlier than female. As the ecog score rises up, the survival probability of patient goes down. What's more, patient with karnofsky score greater than or equal to 70 seems to have higher probability to survive. We will use the Log rank test to confirm our findings.

◆ Log Rank Test

```
survdiff(formula = Surv(time, status) ~ sex, data = data_f2)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 124       99     80.7      4.13      9.25
sex=2  85       48     66.3      5.03      9.25

 Chisq= 9.3  on 1 degrees of freedom, p= 0.002
Call:
survdiff(formula = Surv(time, status) ~ ph.ecog, data = data_f2)
survdiff(formula = Surv(time, status) ~ ph.ecog, data = data_f2)

            N Observed Expected (O-E)^2/E (O-E)^2/V
ph.ecog=0  60       34   49.760   4.99153   7.63221
ph.ecog=1 106       76   75.682   0.00134   0.00277
ph.ecog=2  42       36   21.403   9.95490  11.74539
ph.ecog=3   1        1    0.155   4.61450   4.64658

 Chisq= 19.8  on 3 degrees of freedom, p= 2e-04
survdiff(formula = Surv(time, status) ~ ph.karno, data = data_f2)

             N Observed Expected (O-E)^2/E (O-E)^2/V
ph.karno=1  21       17     13.7    0.7949     0.884
ph.karno=2 188      130    133.3    0.0817     0.884

 Chisq= 0.9  on 1 degrees of freedom, p= 0.3
survdiff(formula = Surv(time, status) ~ pat.karno, data = data_f2)

              N Observed Expected (O-E)^2/E (O-E)^2/V
pat.karno=1  35       30     15.8     12.86      14.6
pat.karno=2 174      117    131.2      1.54      14.6

 Chisq= 14.6  on 1 degrees of freedom, p= 1e-04
survdiff(formula = Surv(time, status) ~ age, data = data_f2)

        N Observed Expected (O-E)^2/E (O-E)^2/V
age=1  86       56       65     1.253      2.26
age=2 123       91       82     0.994      2.26

 Chisq= 2.3  on 1 degrees of freedom, p= 0.1
survdiff(formula = Surv(time, status) ~ wt.loss, data = data_f2)

            N Observed Expected (O-E)^2/E (O-E)^2/V
wt.loss=1 120       78     87.2      0.97      2.41
wt.loss=2  89       69     59.8      1.41      2.41

 Chisq= 2.4  on 1 degrees of freedom, p= 0.1
```

We can find that the results conform to the results of our plots. Sex, ph.ecog and pat.karno have p-values smaller than 0.05 while other variables are not. Hence Sex, ph.ecog and pat.karno are significant in our model.

◆ Backward Elimination

We use the function “step” in R to apply backward elimination method.

```
Step:  AIC=1284.8
Surv(time, status) ~ sex + ph.ecog + ph.karno + pat.karno

            Df  AIC
- ph.karno   1 1284
<none>         1285
- pat.karno  1 1285
- ph.ecog    1 1292
- sex        1 1295

Step:  AIC=1283.9
Surv(time, status) ~ sex + ph.ecog + pat.karno

            Df  AIC
<none>         1284
- pat.karno  1 1286
- ph.ecog    1 1290
- sex        1 1293
```

It indicates that we should include pat.karno, ph.ecog and sex as our variables. We use summary function to further explore the model.

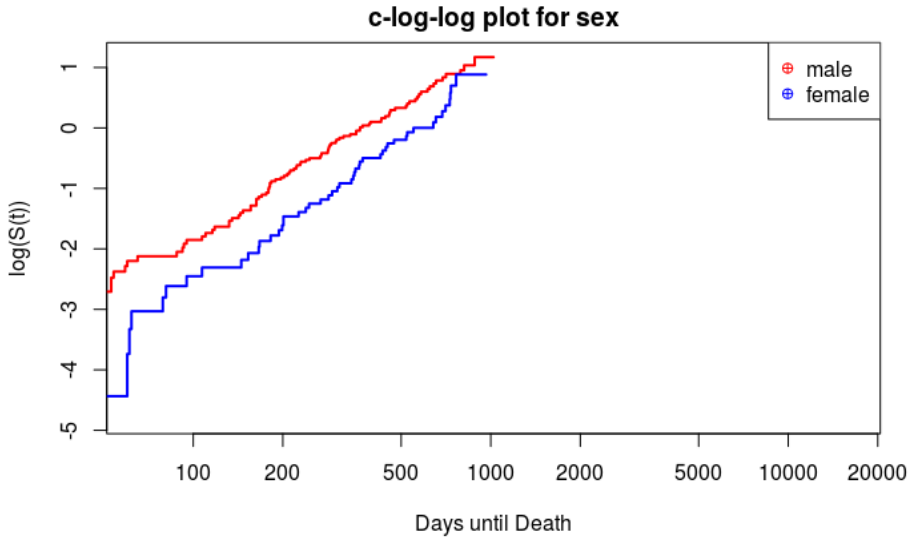
```
> summary(pcox)
Call:
coxph(formula = Surv(time, status) ~ sex + ph.ecog + pat.karno,
    data = data_f2)

  n= 209, number of events= 147

              coef exp(coef) se(coef)     z Pr(>|z|)
sex         -0.575     0.563    0.177 -3.25   0.0011 **
ph.ecog      0.380     1.462    0.134  2.84   0.0045 **
pat.karno2 -0.467     0.627    0.236 -1.97   0.0484 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
sex            0.563      1.777     0.398     0.796
ph.ecog        1.462      0.684     1.125     1.900
pat.karno2     0.627      1.594     0.395     0.997

Concordance= 0.662  (se = 0.026 )
Likelihood ratio test= 30.2  on 3 df,   p=1e-06
Wald test            = 31.8  on 3 df,   p=6e-07
Score (logrank) test = 33  on 3 df,   p=3e-07
```

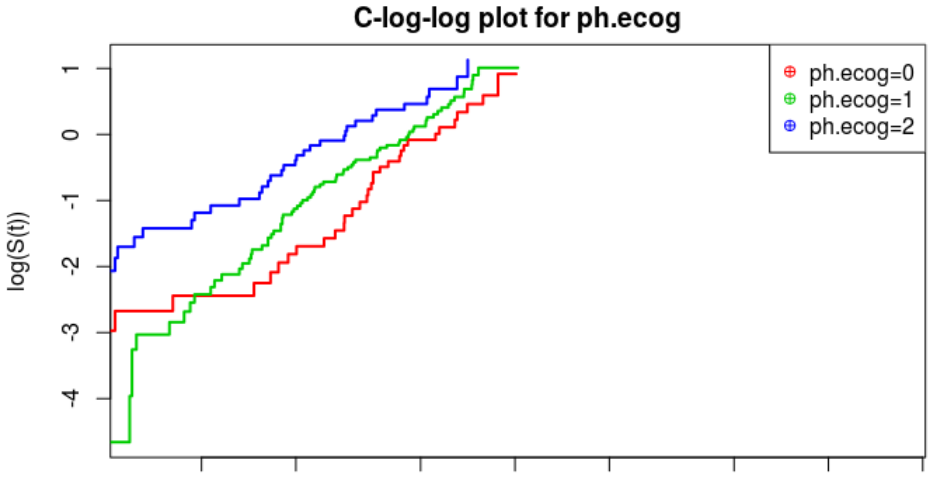
The summary provides us with information that the model is significant in likelihood ratio test, wald test and Score(logrank) test. The coefficients of variables sex,ph.ecog and pat.karno2 are all significant. The coefficient for variables sex2 is negative, implying that female is related with better survival. Patient with a higher score in ph.ecog is related with worse survival since the coefficient is positive.

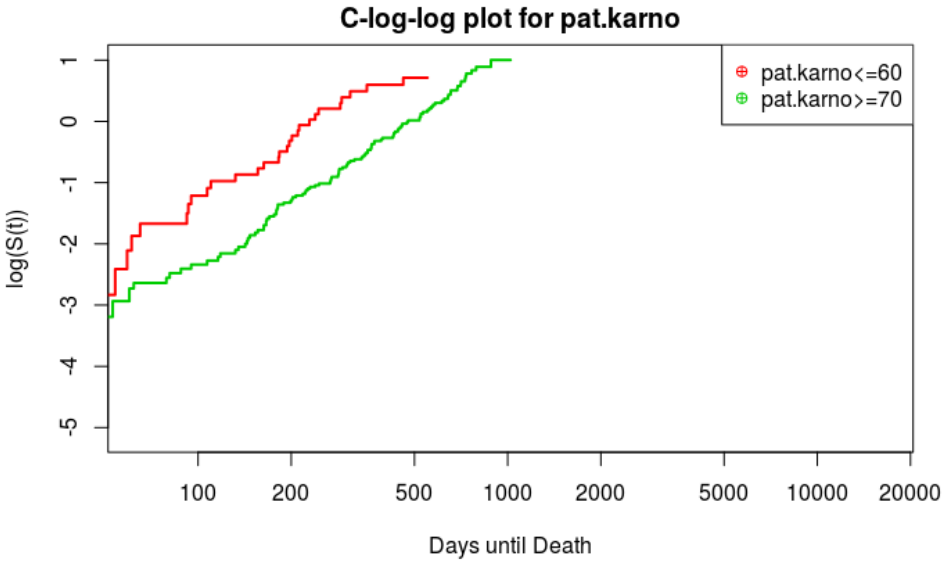
Model Checking

Since we are building the Cox PH model, we need to check whether all the variables satisfy the Ph assumption. We will use both C-log-log plot and residual tests to check Cox Ph assumption.

◆ C-log-log plot


c-log-log plot for sex
male
female
log(S(t))
1
0
-1
-2
-3
-4
-5
100
200
500
1000
2000
5000
10000
20000
Days until Death


C-log-log plot for ph.ecog
ph.ecog=0
ph.ecog=1
ph.ecog=2
log(S(t))
1
0
-1
-2
-3
-4


C-log-log plot for pat.karno
pat.karno<=60
pat.karno>=70
log(S(t))
1
0
-1
-2
-3
-4
-5
100
200
500
1000
2000
5000
10000
20000
Days until Death

From the plot, we see that the lines in the plot for sex and for pat.karno show that these two variables should meet the PH assumption. However, we can see that only in the plot of ph.ecog, there exists a crossing between two lines, which means that ph.ecog doesn't meet the PH assumption. Then we decide to have a more straightforward check for the PH Assumption by the cox.zph test.

```
                rho chisq      p
sex          0.1454  3.04 0.0813
ph.ecog     -0.0808  1.00 0.3165
pat.karno2   0.1039  1.67 0.1968
GLOBAL           NA  7.82 0.0500
```

We find that though all the variables have p value bigger than 0.05, the global's p-value is equal to the 0.05, which reject the null hypothesis that the model is satisfied for the PH assumption. Therefore, since we find that the lines cross in the C-log-log plot of ph.ecog, we decide to stratify the variable ph.ecog since it may conduct to the failure of meeting coxph assumption. Then we check again.

```
              rho chisq     p
sex        0.1264 2.206 0.138
pat.karno2 0.0698 0.825 0.364
GLOBAL         NA 2.946 0.229
```

This time all the variables and Global satisfy the PH assumption and hence, our model becomes Surv ~ sex + pat.karno +strata(ph.ecog).

◆ Interaction term

Now we consider the interaction terms in our model. We have 3 underlying interaction terms which are sex*pat.karno, sex*strata(ph.ecog) and pat.karno*strata(ph.ecog). We will run the likelihood test to check each term.

```
> c(p1,p2,p3)
[1] 0.81302 0.99569 0.41220
```

We find that all of the p-values are bigger than 0.05, hence none of the interaction terms is significant.

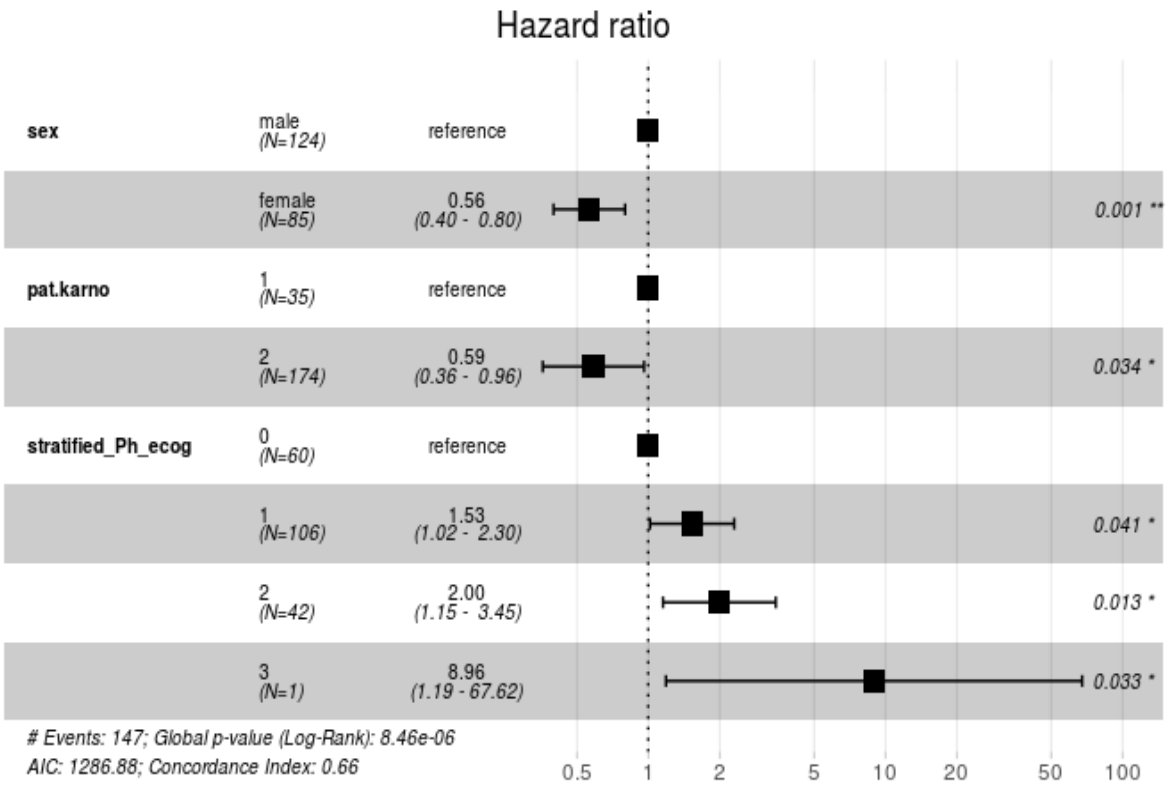
Therefore, our final model should be Surv ~ sex + pat.karno +strata(ph.ecog)

```
> summary(pcox4)
Call:
coxph(formula = Surv(time, status) ~ sex + pat.karno + strata(ph.ecog),
    data = data_f2)

  n= 209, number of events= 147

              coef exp(coef) se(coef)     z Pr(>|z|)
sex         -0.591     0.554    0.180 -3.28   0.0011 **
pat.karno2 -0.536     0.585    0.257 -2.08   0.0371 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
sex            0.554       1.81     0.389     0.789
pat.karno2     0.585       1.71     0.353     0.968

Concordance= 0.609  (se = 0.026 )
Likelihood ratio test= 15  on 2 df,   p=6e-04
Wald test            = 14.7  on 2 df,   p=7e-04
Score (logrank) test = 14.9  on 2 df,   p=6e-04
```

We can see that all the variables are significant and the model pass all the tests (Likelihood ratio test, Wald test and Score (logrank) test)

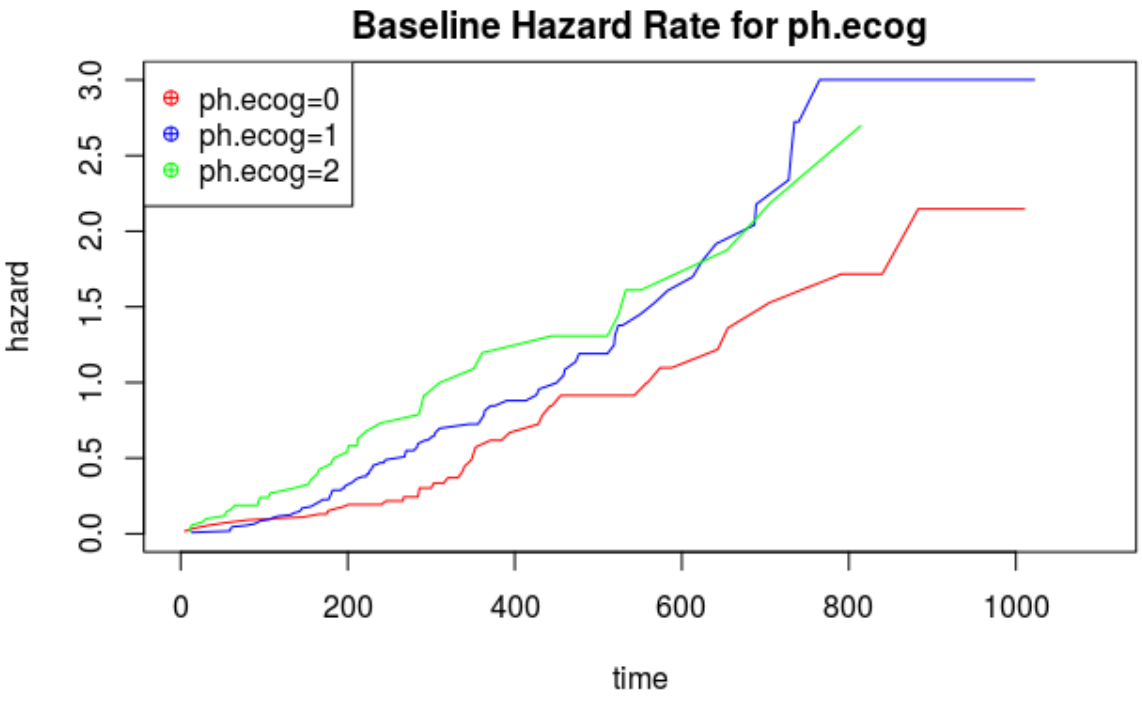
- Hazard Ratios and C.I.

Use ggforest() to find out the Hazard Ratios and C.I. Interval



We can find that the hazard ratio of sex female is centered at 0.56 with 95% confidence interval from 0.4 to 0.8. It means that female are associated with better survival and less than 0.44 (1-0.56) likelihood to die than male due to the lung cancer. For covariate pat.karno, the hazard ratio of pat.karno>=70(coded as 2) is 0.59 compared to pat.karno<=60 (coded as 1). It implies that patient with pat.karno larger or equal to 70 have better chance to survive than those with that smaller or equal to 60. In addition, the 95%CI is (0.36,0.96)

◆ Baseline Hazard Rates



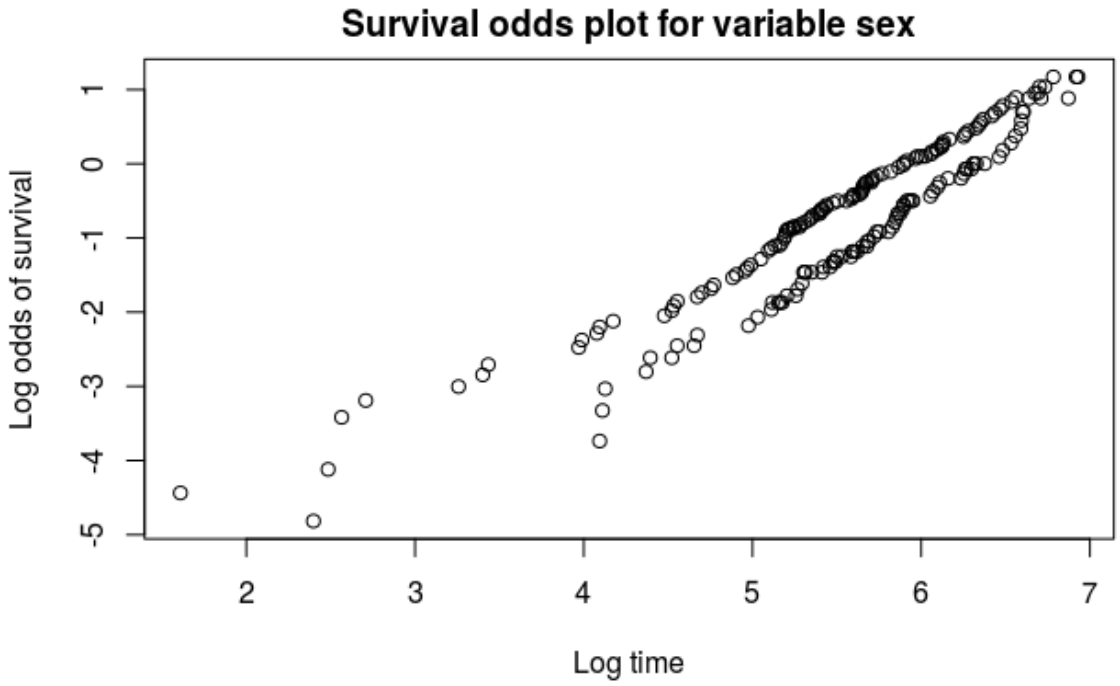
We also draw the baseline hazard plot for each strata, where 0 represents ph.ecoog=0, 1 represents ph.ecog=1 and 2 represents ph.ecog=2. We can easily see that people with ph.ecog equal 0 have the lowest hazard among these 3 groups, implying that they have the highest probability to survive in lung cancer.

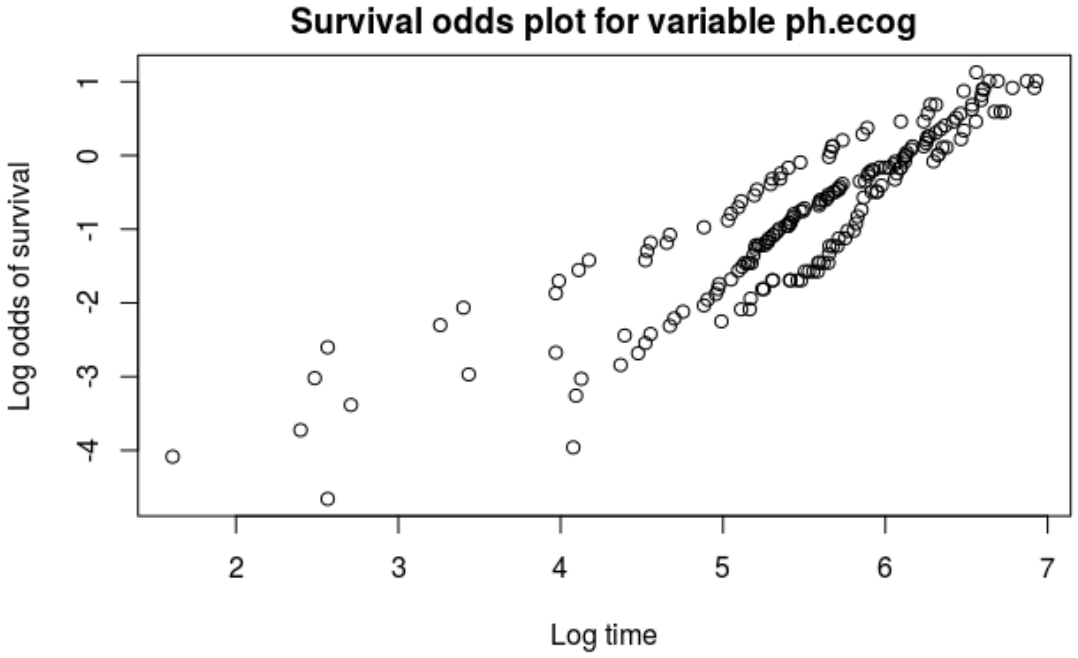
Extension

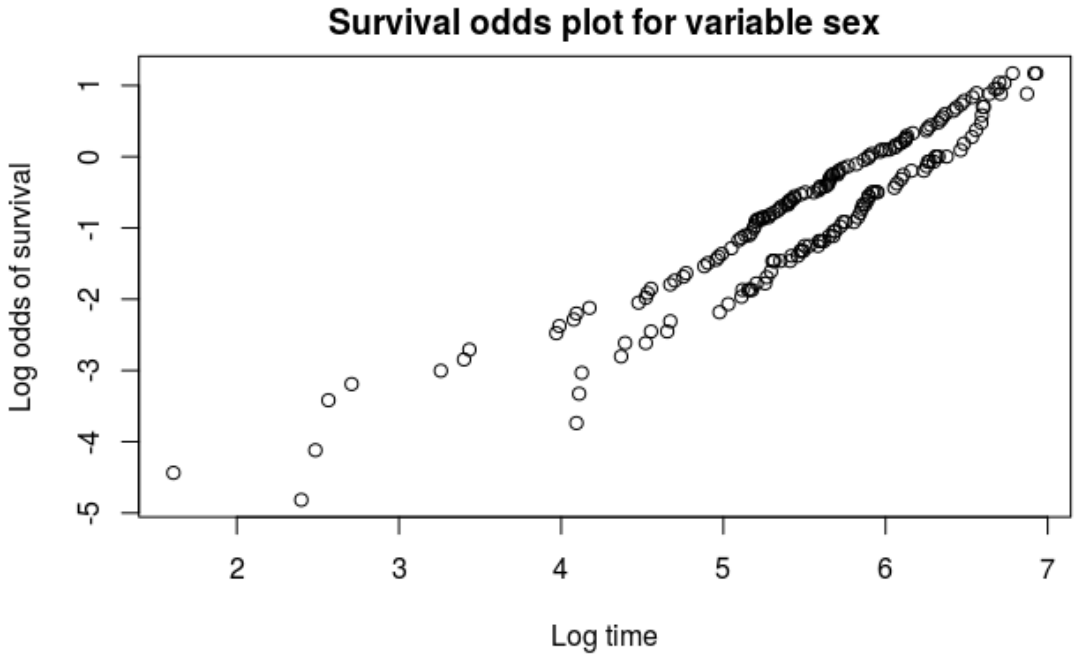
We want to build an AFT model (accelerated failure time model). As an alternative to the cox PH model, it emphasizes the role of the covariates to accelerate or slow down the life course of the disease by some constant. First, we need to decide which distribution is the best among weibull distribution, exponential distribution, Gaussian distribution and logistic distribution.

```
> anova(model_weib,model_exp,model_gaus,model_log)
                        Terms Resid. Df  -2*LL Test Df Deviance
1 sex + pat.karno + ph.ecog       204 2041.8      NA       NA
2 sex + pat.karno + ph.ecog       205 2066.6    = -1 -24.7613
3 sex + pat.karno + ph.ecog       204 2092.2    =  1 -25.5623
4 sex + pat.karno + ph.ecog       204 2089.5    =  0   2.6212
>
```

Since the AIC of model_weib is the smallest, we will choose model_weib as our model. We will check our assumption by drawing the survival odds plot for variables sex, pat.karno and ph.ecog.







From these graphs, the straight lines support weibull distribution. Therefore, we can use Weibull distribution.

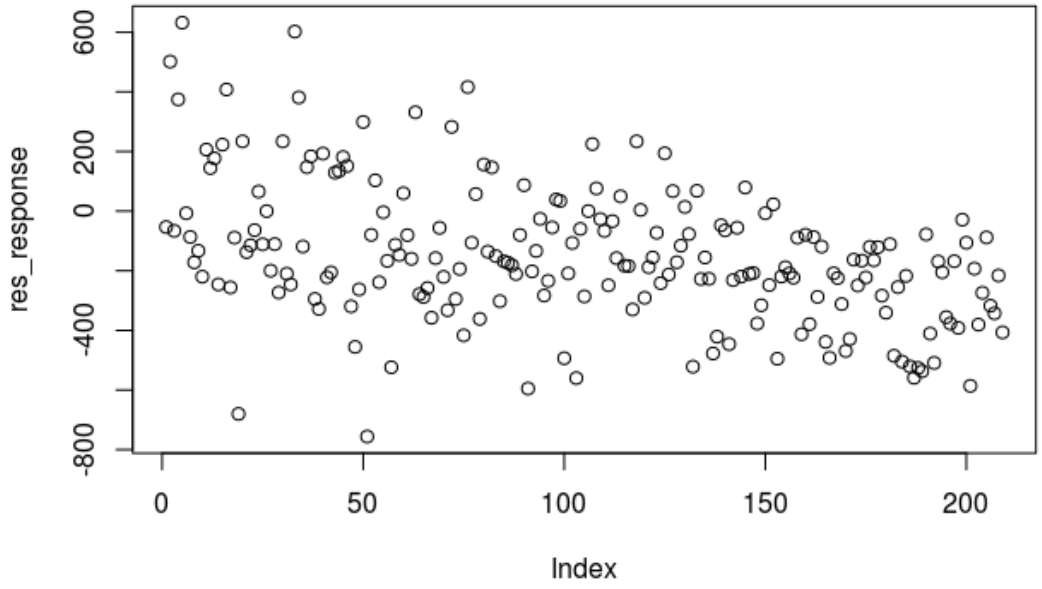
Use summary function to explore the model.

```
Call:
survreg(formula = Surv(time, status) ~ sex + pat.karno + ph.ecog,
    data = data_f2, dist = "weibull")
              Value Std. Error     z       p
(Intercept)  5.4832     0.2615 20.97 < 2e-16
sex          0.4033     0.1261  3.20  0.0014
pat.karno2   0.3450     0.1649  2.09  0.0364
ph.ecog     -0.2653     0.0947 -2.80  0.0051
Log(scale)  -0.3498     0.0643 -5.44 5.3e-08

Scale= 0.705

Weibull distribution
Loglik(model)= -1020.9   Loglik(intercept only)= -1036.1
        Chisq= 30.33 on 3 degrees of freedom, p= 1.2e-06
Number of Newton-Raphson Iterations: 5
n= 209
```

The estimated acceleration factor γ comparing male and female is 1.4968 (e^0.4033). It means that S(male)=S(1.4968*female. Therefore, the time before death for male is “accelerated” by factor of 1.4981 compared to female. It implies that the probability of a male surviving t years equals the probability of a female surviving 1.49033*t years.

In addition, the estimated acceleration factor γ comparing pat.karno=1 (pat.karno<=60) and pat.karno=2 (pat.karno>=70) is 1.412 (e^0.3450). It means that S(pat.karno=1)=S(1.412*pat.karno=2). Therefore, the time before death for patients with karnofsky score lower or equal to 60 is “accelerated” by factor of 1.412 compared to patients with karnofsky score greater or equal to 70. It implies that the probability of a patient with karnofsky score lower or equal to 60 surviving t years equals the probability of a patients with karnofsky score greater or equal to 70 surviving 1.49033*t years.

We also do the residual test for AFT model.



From the residuals plot, we can figure out that the residuals are basically symmetrically distributed around 0. The model is adequate.

Conclusion and Discussion

In this project, we did our analysis on the lung data and tried to figure out how each variable influences the survival time. We first plotted Kaplan-Meier curves and log rank tests to check whether these variables have significant effect on the survival function and we find that sex,pat.karno and ph.ecog have significant influence. We use the backward elimination to figure out our model and then use test cox.zph() and C-log-log plot to check whether the model meet the PH assumption. We find that

ph.ecog and global do not satisfy the PH assumption and hence we stratify the covariate ph.ecog. We then find out the model do not need to include the interaction term. Finally, we summarize the final model by figuring out the hazard ratio, Confidence Interval and Baseline Hazard Function. We can conclude that female have better survival probability than man do and people with higher pat.karno have higher survival probability. We also use AFT model, using weibulll distribution, men or patient with pat.karno lower and equal to 60 have a shorter period to reach the survival probability than people in respectively categories.

However, due to the restricted data, we cannot figure out what influence the difference males and females. Is it due to number of males' smokers are larger than that of females' smokers or other reason? It should be further studied in the other dataset.