Final Project

# Abstract

Lung cancer is one of the most deadly malignancies. Variables such as sex,and ph.ecog,etc. may affect lung cancer patients’ survival time. In this project, we will build a Cox Proportional Hazards model to study the relationship between covariates and survival, using a dataset with 228 patients.

# Data source

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 are as follows:

* inst (Institution code).
* time ( Survival time in days).
* status (censoring status), coded as 1=censored, 2=dead.
* age ( Age in years).
* sex, coed as 1=Male, 2=Female.
* ph.ecog (ECOG performance score), coded as 0=good, 5=dead.
* ph.karno (Karnofsky performance score), coded from 0=bad to 100=good rated by physician.
* pat.karno (Karnofsky performance score as rated by patient).
* meal.cal ( Calories consumed at meals).

library("survival")  
library("survminer")

## Loading required package: ggplot2

## Loading required package: ggpubr

## Loading required package: magrittr

library(tidyverse)

## -- Attaching packages ---------------------------------- tidyverse 1.2.1 --

## √ tibble 1.4.1 √ purrr 0.2.4  
## √ tidyr 0.7.2 √ dplyr 0.7.4  
## √ readr 1.1.1 √ stringr 1.2.0  
## √ tibble 1.4.1 √ forcats 0.2.0

## -- Conflicts ------------------------------------- tidyverse\_conflicts() --  
## x tidyr::extract() masks magrittr::extract()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x purrr::set\_names() masks magrittr::set\_names()

lung <- as\_tibble(lung)  
head(lung)

## # A tibble: 6 x 10  
## inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.l~  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 3.00 306 2.00 74.0 1.00 1.00 90.0 100 1175 NA   
## 2 3.00 455 2.00 68.0 1.00 0 90.0 90.0 1225 15.0  
## 3 3.00 1010 1.00 56.0 1.00 0 90.0 90.0 NA 15.0  
## 4 5.00 210 2.00 57.0 1.00 1.00 90.0 60.0 1150 11.0  
## 5 1.00 883 2.00 60.0 1.00 0 100 90.0 NA 0   
## 6 12.0 1022 1.00 74.0 1.00 1.00 50.0 80.0 513 0

By using the *dim* function we can see that the dataset has 228 obeservations.

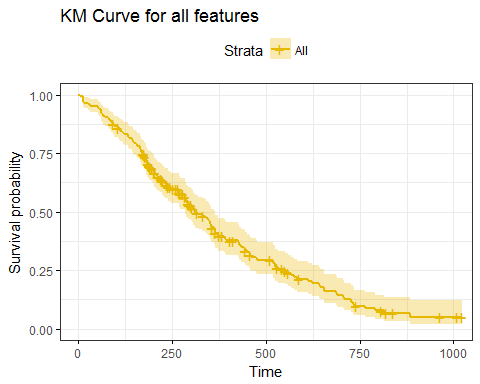
dim(lung)

## [1] 228 10

# Research Question

We are interested in whether age, sex, ph.ecog, ph.karno,pat.karno, meal.cal and wt.loss would lead to diffrence in survival time. And We want to explore the joint effect of multiple variables that may exist.

km0 = survfit(Surv(time,status)~1, data=lung)  
ggsurvplot(km0,  
 ggtheme = theme\_bw(),   
 palette = c("#E7B800", "#2E9FDF"))+ggtitle("KM Curve for all features")



# Data Exploration

## Preprocessing

Before we build the Cox PH model, we try to understand the dataset first. By using the *is.na* function we can get information that variable meal.cal has more than 40 missing values.

apply(lung,2,function(x) sum(sapply(x, is.na)))

## inst time status age sex ph.ecog ph.karno   
## 1 0 0 0 0 1 1   
## pat.karno meal.cal wt.loss   
## 3 47 14

Using 80% rule(Bijlsma et al. 2006)[[1]](#footnote-1), we decide not to consider variable meal.cal in our model and get dataset data\_f0.

miss\_p <- sapply(lung, function(x) sum(is.na(x))/length(x))  
v\_rm <- names(miss\_p)[miss\_p > 1 - 0.8]  
v\_rm

## [1] "meal.cal"

data\_f0=lung[ , !names(lung) %in% c("meal.cal")]

We then drop observations with missing values from data\_f0 and obtain data\_f1. We can get 209 complete observations.

data\_f1=na.omit(data\_f0)  
dim(data\_f1)

## [1] 209 9

## Description

Before doing data description, we change the type of variables status,sex,ph.ecog,ph.karno and pat.karno from numeric to factor.

cols=c('sex','ph.ecog','ph.karno','pat.karno')  
data\_f1[,cols] <- lapply(data\_f1[,cols],as.factor)

By using the *summary* function, we find that the age of people is distributed between 39 and 82 years old, 124 out of 209 observations are male, and 166 out of 209 observations get ECOG performance score under 2, only 29 of the obsevations scored 100 in Karnofsky performance, the pat.karno of people is distributed between 30 and 100, and the weight loss of no less than 50% of people is 6.

summary(data\_f1)

## inst time status age sex   
## Min. : 1.00 Min. : 5.0 Min. :1.000 Min. :39.00 1:124   
## 1st Qu.: 3.00 1st Qu.: 176.0 1st Qu.:1.000 1st Qu.:56.00 2: 85   
## Median :11.00 Median : 269.0 Median :2.000 Median :63.00   
## Mean :10.91 Mean : 317.7 Mean :1.703 Mean :62.38   
## 3rd Qu.:16.00 3rd Qu.: 429.0 3rd Qu.:2.000 3rd Qu.:70.00   
## Max. :33.00 Max. :1022.0 Max. :2.000 Max. :82.00   
##   
## ph.ecog ph.karno pat.karno wt.loss   
## 0: 60 50 : 4 90 :59 Min. :-24.000   
## 1:106 60 :17 80 :47 1st Qu.: 0.000   
## 2: 42 70 :29 70 :36 Median : 6.000   
## 3: 1 80 :63 100 :32 Mean : 9.584   
## 90 :67 60 :28 3rd Qu.: 15.000   
## 100:29 50 : 3 Max. : 68.000   
## (Other): 4

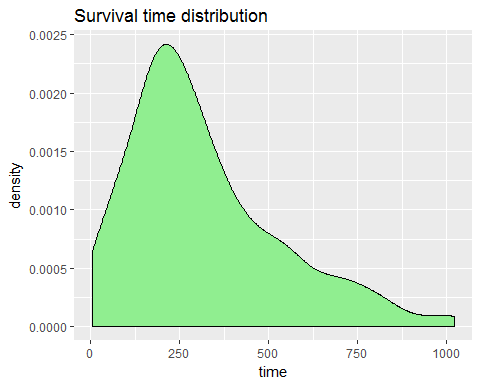
By applying the function *quantile* in R, we see that 62.5% of the observations survive less than 320 days.

quantile(data\_f1$time,probs=seq(0,1,by=0.125))

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

We also plot the survival time distribution, it provides more information for the survival time of the observations.

ggplot(data\_f1, aes(x = time)) + geom\_density(fill = "light green") + ggtitle("Survival time distribution")



# Model building

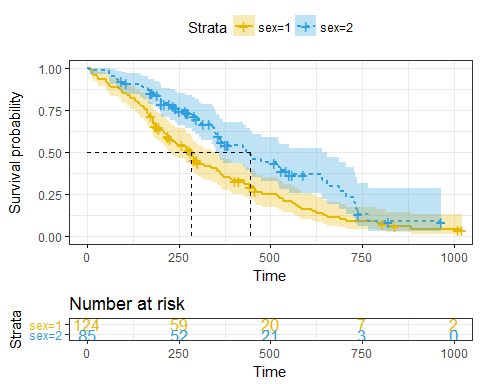
Now, we start to build our Cox PH model using the following variables: age, sex, ph.ecog, ph.karno, pat.karno and wt.loss.

## Kaplan-Mayer estimation and log-rank test

We try to plot the Kaplan-Meier survival curves to visually analyze the effects of each catagorical covariate sex,ph.ecog ,ph.karno and pat.karno on survival. For each of the group, we perform Kaplan-Mayer estimation and log-rank test.

### Covariate sex

km1 = survfit(Surv(time,status)~sex, data=data\_f1)  
ggsurvplot(km1,  
 conf.int = TRUE,  
 risk.table = TRUE,   
 risk.table.col = "strata",   
 linetype = "strata",   
 surv.median.line = "hv",   
 ggtheme = theme\_bw(),   
 palette = c("#E7B800", "#2E9FDF"))



From the output, we can see there is a significant difference between male and female.Then we perform the log-rank test:

survdiff(Surv(time,status)~sex, data=data\_f1)

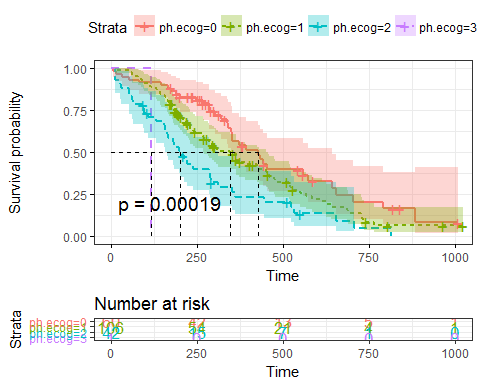
## Call:  
## survdiff(formula = Surv(time, status) ~ sex, data = data\_f1)  
##   
## N Observed Expected (O-E)^2/E (O-E)^2/V  
## sex=1 124 99 80.7 4.13 9.25  
## sex=2 85 48 66.3 5.03 9.25  
##   
## Chisq= 9.3 on 1 degrees of freedom, p= 0.00235

The p-value is 0.00235, which supports to reject the null hypothesis.

### Covariate ph.ecog

The p-value can also be shown in the graph. For covariate ph.ecog, we could see both the plot and the log-rank p-value suggest that it is significant for difference of the survival functions.

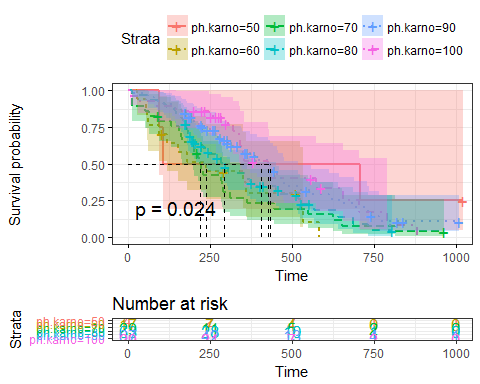
km2 = survfit(Surv(time,status)~ph.ecog, data=data\_f1)  
ggsurvplot(km2,  
 pval = TRUE, conf.int = TRUE,  
 risk.table = TRUE,   
 risk.table.col = "strata",  
 linetype = "strata",   
 surv.median.line = "hv",   
 ggtheme = theme\_bw(),   
 )



### Covariate ph.karno

With the plot and the log-rank p-value smaller than 0.05, we could reject the null hypothesis that the groups have the same survival function.

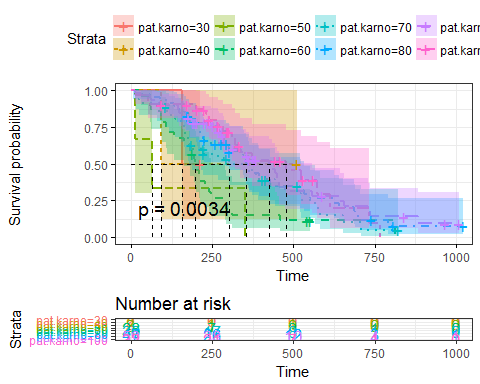
km3 = survfit(Surv(time,status)~ph.karno, data=data\_f1)  
ggsurvplot(km3,  
 pval = TRUE, conf.int = TRUE,  
 risk.table = TRUE,   
 risk.table.col = "strata",   
 linetype = "strata",   
 surv.median.line = "hv",   
 ggtheme = theme\_bw()  
 )



### Covariate pat.karno

As the p-value is 0.0034, it is acceptable to say that pat.karno is significantly different for survival functions.

km4 = survfit(Surv(time,status)~pat.karno, data=data\_f1)  
ggsurvplot(km4,  
 pval = TRUE, conf.int = TRUE,  
 risk.table = TRUE,   
 risk.table.col = "strata",   
 linetype = "strata",   
 surv.median.line = "hv",   
 ggtheme = theme\_bw()  
 )



## Univariate analysis

Since Kaplan-Meier curves and log-rank tests are useful only when the predictor variable is categorical, we apply Cox proportional hazards regression analysis to the continious variables.

covariates<-c("age","wt.loss")  
td\_formulas<-sapply(covariates,function(x)  
 as.formula(paste('Surv(time, status)~',x)))  
td\_models<-lapply(td\_formulas,function(x){coxph(x,data=data\_f1)})  
td\_results <- lapply(td\_models,  
 function(x){   
 x <- summary(x)  
 p.value<-signif(x$wald["pvalue"], digits=2)  
 wald.test<-signif(x$wald["test"], digits=2)  
 beta<-signif(x$coef[1], digits=2);  
 HR <-signif(x$coef[2], digits=2);  
 HR.confint.lower <- signif(x$conf.int[,"lower .95"],2)  
 HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)  
 HR <- paste0(HR, " (",   
 HR.confint.lower, "-", HR.confint.upper, ")")  
 res<-c(beta, HR, wald.test, p.value)  
 names(res)<-c("beta", "HR (95% CI for HR)", "wald.test", "p.value")  
 return(res)  
   
 })  
results\_a <- t(as.data.frame(td\_results, check.names = FALSE))  
as.data.frame(results\_a)

## beta HR (95% CI for HR) wald.test p.value  
## age 0.02 1 (1-1) 4.4 0.036  
## wt.loss -0.00017 1 (0.99-1) 0 0.98

From the output above, we figure out two conclusions: 1.The variables age is statistically significant, while wt.loss does not have statistically significant coefficients.So in the Multivariate Cox regression analysis below, we will not take variable wt.loss into considerration. 2.The coefficient for variables age is positive. Therefore, older age is related with worse survival.

# Multivariate Cox regression analysis

We then try to build a time-independent cox proportional hazard model to see jointly impact on survival of the variables, and we need to decide how many variables should be included into the model. Firstly We build a full model with all five covariates(sex, age, ph.ecog, ph.karno, pat.karno), and use backward elimination to complete variable selection. Model checking will also be done in this part.

## Full model

mul.cox <- coxph(Surv(time, status) ~ age + sex + ph.ecog+ph.karno+pat.karno, data =data\_f1)  
summary(mul.cox)

## Call:  
## coxph(formula = Surv(time, status) ~ age + sex + ph.ecog + ph.karno +   
## pat.karno, data = data\_f1)  
##   
## n= 209, number of events= 147   
##   
## coef exp(coef) se(coef) z Pr(>|z|)   
## age 0.01444 1.01454 0.01072 1.347 0.177874   
## sex2 -0.63579 0.52952 0.18643 -3.410 0.000649 \*\*\*  
## ph.ecog1 0.48497 1.62413 0.30326 1.599 0.109770   
## ph.ecog2 0.91648 2.50046 0.46897 1.954 0.050672 .   
## ph.ecog3 2.25708 9.55516 1.14660 1.969 0.049010 \*   
## ph.karno60 1.02612 2.79023 0.68035 1.508 0.131493   
## ph.karno70 1.05304 2.86634 0.64197 1.640 0.100939   
## ph.karno80 1.30598 3.69131 0.64666 2.020 0.043427 \*   
## ph.karno90 1.17402 3.23496 0.66194 1.774 0.076128 .   
## ph.karno100 1.31848 3.73772 0.72896 1.809 0.070498 .   
## pat.karno40 -0.67877 0.50724 1.47686 -0.460 0.645802   
## pat.karno50 0.73238 2.08002 1.21032 0.605 0.545106   
## pat.karno60 -0.07403 0.92865 1.03472 -0.072 0.942965   
## pat.karno70 -0.41387 0.66109 1.05567 -0.392 0.695025   
## pat.karno80 -0.37357 0.68827 1.06298 -0.351 0.725258   
## pat.karno90 -0.46989 0.62507 1.06663 -0.441 0.659548   
## pat.karno100 -0.59321 0.55255 1.07912 -0.550 0.582513   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## age 1.0145 0.9857 0.99346 1.0361  
## sex2 0.5295 1.8885 0.36744 0.7631  
## ph.ecog1 1.6241 0.6157 0.89638 2.9428  
## ph.ecog2 2.5005 0.3999 0.99732 6.2691  
## ph.ecog3 9.5552 0.1047 1.00984 90.4112  
## ph.karno60 2.7902 0.3584 0.73541 10.5865  
## ph.karno70 2.8663 0.3489 0.81448 10.0873  
## ph.karno80 3.6913 0.2709 1.03930 13.1105  
## ph.karno90 3.2350 0.3091 0.88395 11.8389  
## ph.karno100 3.7377 0.2675 0.89560 15.5992  
## pat.karno40 0.5072 1.9714 0.02806 9.1690  
## pat.karno50 2.0800 0.4808 0.19402 22.2996  
## pat.karno60 0.9286 1.0768 0.12221 7.0568  
## pat.karno70 0.6611 1.5127 0.08350 5.2342  
## pat.karno80 0.6883 1.4529 0.08569 5.5280  
## pat.karno90 0.6251 1.5998 0.07727 5.0565  
## pat.karno100 0.5526 1.8098 0.06665 4.5806  
##   
## Concordance= 0.665 (se = 0.028 )  
## Rsquare= 0.175 (max possible= 0.998 )  
## Likelihood ratio test= 40.22 on 17 df, p=0.001204  
## Wald test = 42.58 on 17 df, p=0.0005524  
## Score (logrank) test = 46.94 on 17 df, p=0.0001245

The full model passes all three overall tests (likelihood, Wald, and score). But there are some variables in the model fail to be significant, so we need to apply feature selection to make the model better.

## Backward elimination

The AIC procedure suggests us a model with variables sex and ph.ecog.

mul.cox1 = step(mul.cox,direction="backward")

## Start: AIC=1301.88  
## Surv(time, status) ~ age + sex + ph.ecog + ph.karno + pat.karno  
##   
## Df AIC  
## - pat.karno 7 1292.4  
## - ph.karno 5 1297.4  
## - ph.ecog 3 1301.3  
## - age 1 1301.7  
## <none> 1301.9  
## - sex 1 1312.2  
##   
## Step: AIC=1292.39  
## Surv(time, status) ~ age + sex + ph.ecog + ph.karno  
##   
## Df AIC  
## - ph.karno 5 1289.7  
## <none> 1292.4  
## - age 1 1292.7  
## - ph.ecog 3 1294.6  
## - sex 1 1303.9  
##   
## Step: AIC=1289.67  
## Surv(time, status) ~ age + sex + ph.ecog  
##   
## Df AIC  
## - age 1 1289.2  
## <none> 1289.7  
## - sex 1 1298.4  
## - ph.ecog 3 1298.8  
##   
## Step: AIC=1289.17  
## Surv(time, status) ~ sex + ph.ecog  
##   
## Df AIC  
## <none> 1289.2  
## - sex 1 1297.9  
## - ph.ecog 3 1300.6

Our final cox model would be:

summary(mul.cox1)

## Call:  
## coxph(formula = Surv(time, status) ~ sex + ph.ecog, data = data\_f1)  
##   
## n= 209, number of events= 147   
##   
## coef exp(coef) se(coef) z Pr(>|z|)   
## sex2 -0.5671 0.5672 0.1774 -3.196 0.00139 \*\*   
## ph.ecog1 0.4533 1.5735 0.2080 2.179 0.02932 \*   
## ph.ecog2 0.9619 2.6166 0.2409 3.993 6.52e-05 \*\*\*  
## ph.ecog3 2.1539 8.6187 1.0305 2.090 0.03661 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## sex2 0.5672 1.7631 0.4006 0.8031  
## ph.ecog1 1.5735 0.6355 1.0466 2.3656  
## ph.ecog2 2.6166 0.3822 1.6320 4.1954  
## ph.ecog3 8.6187 0.1160 1.1435 64.9575  
##   
## Concordance= 0.647 (se = 0.027 )  
## Rsquare= 0.121 (max possible= 0.998 )  
## Likelihood ratio test= 26.93 on 4 df, p=2.05e-05  
## Wald test = 28.15 on 4 df, p=1.166e-05  
## Score (logrank) test = 30.29 on 4 df, p=4.265e-06

## Model checking

### Scaled Schoenfeld Residuals

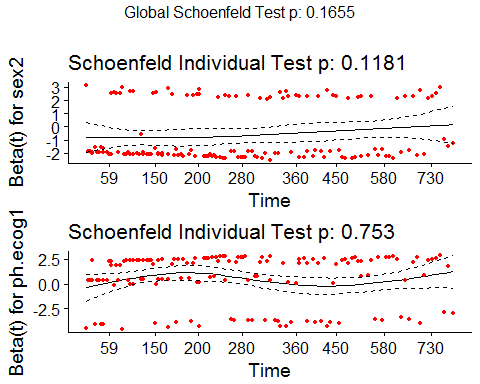
Then we perform PH assumption test on independence between Schoenfeld residual and time. It could be obtained using function *cox.zph()* in R.The P values of the covariates are greater than 0.05, indicating that each variable satisfies the PH assumption check. And the overall P value of the model is not statistically significant. Thus, the model is suitable.

test1.model <- cox.zph(mul.cox1)  
test1.model

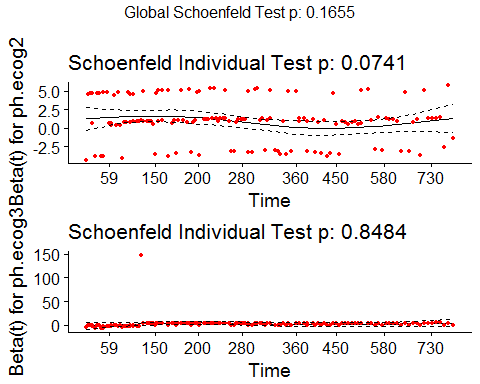
## rho chisq p  
## sex2 0.1306 2.4418 0.1181  
## ph.ecog1 -0.0261 0.0990 0.7530  
## ph.ecog2 -0.1500 3.1890 0.0741  
## ph.ecog3 0.0158 0.0366 0.8484  
## GLOBAL NA 6.4880 0.1655

Use function *ggcoxzph()* in R we can also do diagnostic in a graphical way.

out1=test1.model  
ggcoxzph(out1, var = colnames(out1$y)[1:2])



ggcoxzph(out1, var = colnames(out1$y)[3:4])



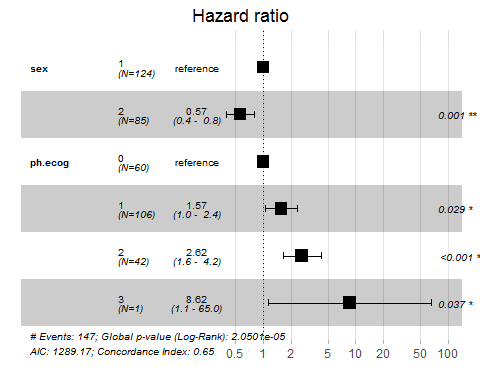
From the figures above, we see that the solid line is a smoothing spline fit to the plot, and the dashed lines represents a 2 times standard-error band around the fit. The graphical diagnostic shows schoenfeld residuals are independent of time.

### Hazard ratios and C.I

Function *ggforest* helps us to find the confidence interval of the hazard ratio.

ggforest(mul.cox1,data=data\_f1)

## Warning: Removed 2 rows containing missing values (geom\_errorbar).

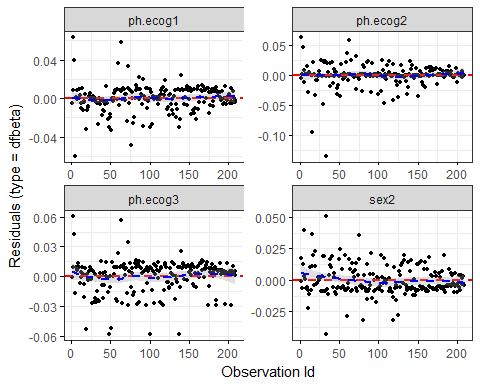


The hazard ratio of sex female is centered at 0.57 with 95% confidence interval 0.4 to 0 .8. It means being female (sex = 2) are associated with better survival. For covariate ph.ecog, we could see that the higher the ph.ecog, the higher the hazard ratio.

### Check influential observations

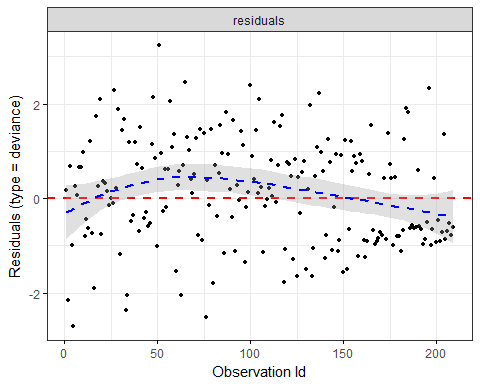
The function *ggcoxdiagnostics()* provides us a solution for finding out influential observations.

ggcoxdiagnostics(mul.cox1, type = "dfbeta",  
 linear.predictions = FALSE, ggtheme = theme\_bw())



From the output above, we can get the conclusion that even though some dfbeta values for the covariates are large comparing with other values, these points have no decisive effect on the overall observations individually.

ggcoxdiagnostics(mul.cox1, type = "deviance",  
 linear.predictions = FALSE, ggtheme = theme\_bw())



Also, from the deviance residuals, we can see the residuals are roughtly symmetrically distributed around zero, showing a good pattern.

## Interaction term

Now, we take interaction terms into consideration. The potential interaction term for our model is: sex\*ph.ecog. According to the univariate analysis, the interaction term is not significant, it fails to pass the likelihood test.Therefore, our final model is Surv~sex+ph.ecog.

data\_add=as.numeric(data\_f1$sex)\*as.numeric(data\_f1$ph.ecog)  
data\_f2=cbind(data\_f1,data\_add)  
mul.cox2 <- coxph(Surv(time, status) ~ data\_add, data =data\_f2)  
summary(mul.cox2)

## Call:  
## coxph(formula = Surv(time, status) ~ data\_add, data = data\_f2)  
##   
## n= 209, number of events= 147   
##   
## coef exp(coef) se(coef) z Pr(>|z|)  
## data\_add 0.03209 1.03261 0.05792 0.554 0.58  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## data\_add 1.033 0.9684 0.9218 1.157  
##   
## Concordance= 0.523 (se = 0.026 )  
## Rsquare= 0.001 (max possible= 0.998 )  
## Likelihood ratio test= 0.3 on 1 df, p=0.5816  
## Wald test = 0.31 on 1 df, p=0.5796  
## Score (logrank) test = 0.31 on 1 df, p=0.5795

# AFT Model

For extension, we try to build a AFT model. First,we need to decide which of the distributions is better.

model\_weib <- survreg(Surv(time, status) ~ sex + ph.ecog ,data=data\_f1,na.action=na.omit, dist="weibull")  
model\_exp <- survreg(Surv(time, status) ~ sex + ph.ecog ,data=data\_f1,na.action=na.omit, dist="exponential")  
model\_gaus <- survreg(Surv(time, status) ~ sex + ph.ecog ,data=data\_f1,na.action=na.omit, dist="gaussian")  
model\_log <- survreg(Surv(time,status)~sex + ph.ecog ,data=data\_f1,na.action=na.omit, dist="logistic")  
result= anova(model\_weib,model\_exp,model\_gaus,model\_log)[,3]  
names(result) = c("model\_weib", "model\_exp", "model\_gaus", "model\_log")  
result

## model\_weib model\_exp model\_gaus model\_log   
## 2045.584 2069.120 2096.507 2094.804

Since the AIC of model\_weib is the smallest, we will choose model\_weib as our model.

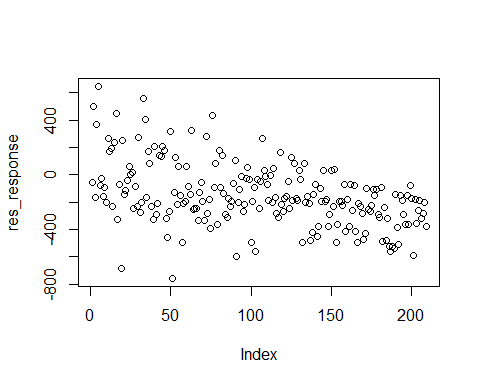
summary(model\_weib)

##   
## Call:  
## survreg(formula = Surv(time, status) ~ sex + ph.ecog, data = data\_f1,   
## na.action = na.omit, dist = "weibull")  
## Value Std. Error z p  
## (Intercept) 6.237 0.1260 49.48 0.00e+00  
## sex2 0.399 0.1275 3.13 1.75e-03  
## ph.ecog1 -0.312 0.1475 -2.12 3.43e-02  
## ph.ecog2 -0.684 0.1729 -3.95 7.67e-05  
## ph.ecog3 -1.467 0.7213 -2.03 4.20e-02  
## Log(scale) -0.342 0.0648 -5.28 1.27e-07  
##   
## Scale= 0.71   
##   
## Weibull distribution  
## Loglik(model)= -1022.8 Loglik(intercept only)= -1036.1  
## Chisq= 26.59 on 4 degrees of freedom, p= 2.4e-05   
## Number of Newton-Raphson Iterations: 5   
## n= 209

The estimated acclerated factor comparing sex female with sex male is 1.49033(e^0.399), which means that the probability of a male surviving t years equals the probability of a female surviving 1.49033\*t years. And likewise, ph. estimated acclerated factor comparing ph.ecog2 with ph.ecog0 is 0.50459(e^-0.684), which means that the probability of an observation scored 2 in ph.ecog surviving t years equals the probability of an observation scored 0 in ph.eocg surviving 0.50459\*t years.

## Residuals

res\_response=residuals(model\_weib, type="response")  
plot(res\_response)



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

# Conclusion and discussion

In this project, We did our analysis on the lung data and tried to figure out how each variable affects the survival time.We analyzed the relationship between single categorical variables and survival through Kaplan-Meier curves and logrank test tests. For numerical variables,we used single factor cox regression analysis to explore the relationship between a single numerical variable and survival.In order to fit the multi-factor cox proportional hazards model, we applied the AIC criterion to select the appropriate model, and then tested the model hypotheses, and found that all variables and the model met the hypotheses. We then tested the possible interaction term and found it had no significant effect on survival time.

In the expansion phase, we tried AFT models, and compared AFT models corresponding to different distributions, and selected the Weibull model according to AIC criteria.

1. Bijlsma, Sabina, Ivana Bobeldijk, Elwin R. Verheij, Raymond Ramaker,Sunil Kochhar, Ian A. Macdonald, Ben Van Ommen, and Age K. Smilde. 2006.Large-scale human metabolomics studies: A strategy for data (pre-)processing and validation. *Analytical Chemistry* 78 (2):567–74. [↑](#footnote-ref-1)