First Assignment - Biostatistics

```
#Loading packages
library(survminer)
library(survival)
```

1. Consider a survival function with constant hazard h(t) = 0,07 when 0 ≤ t ≤ 5, and h(t) = 0,4 for t > 5 (This is known as a piecewise constant hazard.) Plot this hazard function and the corresponding survival function for 0 < t < 10. What is the median survival time?

First of all, I calculate the vector of h(t) for each unit of time and then, we apply the relation: s(t)=h(t)*t in order to calculate the vector of s(t).

```
ht1=0.07
ht2=0.4

ht=c(0)
for (i in 1:10) {
  if (i<6) {
    ht[i+1]=ht1
  } else {
    ht[i+1]=ht2
  }
}
ht
```

```
##  [1] 0.00 0.07 0.07 0.07 0.07 0.07 0.40 0.40 0.40 0.40 0.40
```

```
t=c(0,1,2,3,4,5,6,7,8,9,10)
st=exp(-ht*t)
st
```

```
##  [1] 1.00000000 0.93239382 0.86935824 0.81058425 0.75578374 0.70468809
##  [7] 0.09071795 0.06081006 0.04076220 0.02732372 0.01831564
```

Then, we plot both vector versus time, obtaning the figure 1 and figure 2.

```
hazard <- plot(x=t,y=ht,type='s')
```

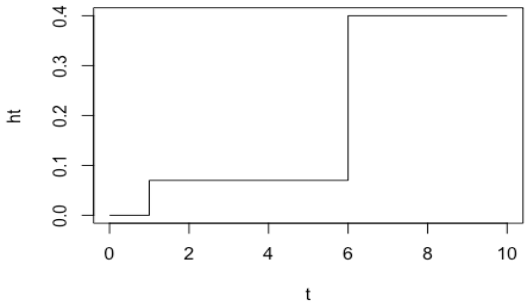

Figure 1. Hazard function

```
survival <- plot(x=t,y=st,type='s')
```

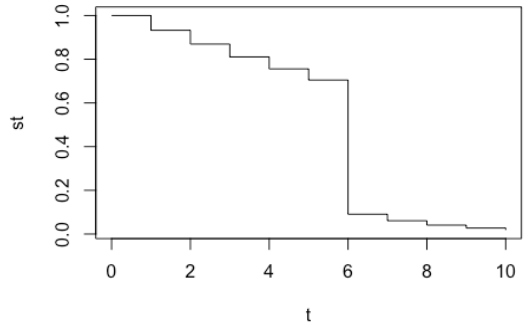

Figure 2. Survival function

Finally, we obtain the median.

```
median(st)
## [1] 0.7046881
```

The median of the survival time is the middle value, and therefore, it is 0.705.

2. Suppose that we assume that the time-to-even is a Rayleigh distribution with density function:

$$f(y) = (\lambda_0 + \lambda_1 y)\exp\left(-\lambda_0 y - \frac{1}{2}\lambda_1 y^2\right), y > 0$$

Calculate the survival, hazard and cummulative hazard functions.

- Survival function (s(y)):

$$S(y) = \int_0^{\infty} f\,(y)\partial y$$

$$S(y) = \int_0^{\infty} (\lambda_0 + \lambda_1 y)\exp\left(-\lambda_0 y - \frac{1}{2}\lambda_1 y^2\right)\partial y$$

$$\boldsymbol{S(y) = -\exp\left(-\frac{1}{2}\lambda_1 y^2 - \lambda_0 y\right)\ \text{where}\ y, \lambda_0, \lambda_1 > 0}$$

- Hazard function (h(y)):

$$h(y) = \frac{f(y)}{s(y)} = \frac{(\lambda_0 + \lambda_1, y)\exp\left(-\lambda_0 y - \frac{1}{2}\lambda_1 y^2\right)}{-\exp\left(-\frac{1}{2}\lambda_1 y^2 - \lambda_0 y\right)}$$

$$\boldsymbol{h(y) = -\lambda_0 - \lambda_1 y}$$

- Cummulative hazard function (H(y)):

$$H(y) = -\log(s(y))$$

$$H(y) = -\log\left(-\exp\left(-\frac{1}{2}\lambda_1 y^2 - \lambda_0 y\right)\right) \quad \text{if } y, \lambda_0, \lambda_1 > 0$$

Then,

$$\boldsymbol{H(y) = \frac{1}{2}\lambda_1 y^2 + \lambda_0 y}$$

3. Construct your own function to obtain the KM estimator of the survival function. Use it to obtaine the survival function of the leukemia data used in the course notes. Plot the function.

Initially, I load the dataset of the AML.

```
aml <- aml
```

Then, I define my own function of KM estimator for survival analysis.

```
MyKM <- function (data) {

  vec=c()

  #Only taking into account the subjects with treatment and distinguish the cens
ored data
  for (i in 1:length(data[,3])) {
    if (data[i,3]=='Maintained') {
      vec[i]=data[i,1]
    }
    if (data[i,3]=='Maintained' && data[i,2]==0) {
      vec[i]='NA'
    }
  }

  #creating the variables of nj,dj and sj
  nj=c(length(vec))
  dj=1
  sj=c(1)

  #Applying the formula (1-(dj/nj)) in order to obtain the survival value per ea
ch single time event
  for (j in 1:(length(vec))) {

    #Control loop to not do do so if the time event is censored
    if (vec[j]=='NA') {
      nj[j+1]=length(vec)-(j-1)
      sj[j+1]=sj[j]
      next
    } else {
      nj[j+1]=length(vec)-(j-1)
      sj[j+1]=(1-(dj/nj[j+1]))*sj[j]
    }
  }

  #converting the outputs to the real needed info
  nj[1]=0
  a=which(vec == 'NA')
  nj=(nj[-c(a)])
  sj=sj[-c(a)]
  print(list(sort(nj),sj))
  plot(x=sort(nj),y=sj, xlab="Time (weeks)",ylab="Survival S(t)",type='s')
}
```

Finally, I apply my function to the dataset 'aml' and plot the result of the function (figure 3).

```
MyKM(aml)

## [[1]]
## [1]  0  1  3  5  6  8  9 11
##
## [[2]]
## [1] 1.0000000 0.9090909 0.8181818 0.7159091 0.6136364 0.4909091 0.3681818
## [8] 0.1840909
```

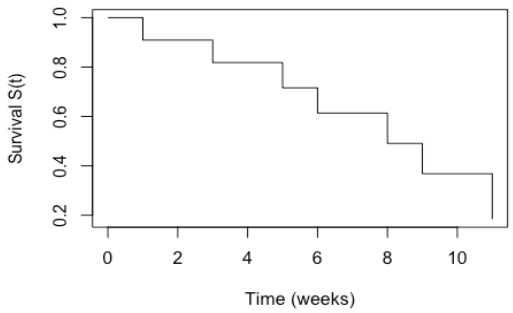

Figure 3. Survival function of MyKM

4. The file 'Henning.txt' contains data from yet another study of criminal recidivism, this one by Henning and Frueh (1996), who followed 194 inmates released from a mediumsecurity prison to a maximum of three years from the day of their release; during the period of the study, 106 of the released prisoners were rearrested. The data set contains the following variables:

- *months: The time of re-arrest in months (but measured to the nearest day).*
- *censor: A dummy variable coded 1 for censored observations and 0 for uncensored observations. Note that this is the opposite of our usual convention!*
- *personal: A dummy variable coded 1 for prisoners with a record of crime against persons and 0 otherwise.*
- *property: A dummy variable coded 1 for prisoners with a record of crime against property and otherwise.*
- *Cage: Centered age in years at time of release, that is, age-average age.*

a) Compute and graph the Kaplan-Meier estimate of the survival function for all of the data.

First of all, I load the dataset and separate in dummy variables the kind of the commited crime.

Also, we change the type of some variables in order to have categorical data instead of numeric and I inverse the values for the censored data since this variables has been completed in the opposite way.

Finally, I apply the function 'survfit' in order to obtain the survival function and I plot it (figure 4).

```
hen <- read.table('Henning.txt', header=T)

for (i in 1:length(hen[,4])) {
  if (hen[i,4]=='1' & hen[i,5]=='1') {
    hen[i,7]="Persons and Property"
  } else if (hen[i,4]=='1' & hen[i,5]=='0') {
      hen[i,7]="Only Persons"
  } else if (hen[i,4]=='0' & hen[i,5]=='1') {
    hen[i,7]="Only Property"
  } else if (hen[i,4]=='0' & hen[i,5]=='0') {
    hen[i,7]="Other crime"
  }
}

hen$personal <- factor(hen$personal,
                       levels = c(0, 1),
                       labels = c("no", "yes"))
hen$property <- factor(hen$property,
                       levels = c(0,1),
                       labels = c("no", "yes"))

hen$censor=ifelse(hen$censor==1,0,1)
Surv(hen$months, hen$censor)

fit1=survfit(Surv(hen$months, hen$censor)~1)
summary(fit1)

plot(fit1, xlab="Time",ylab="Survival")
```

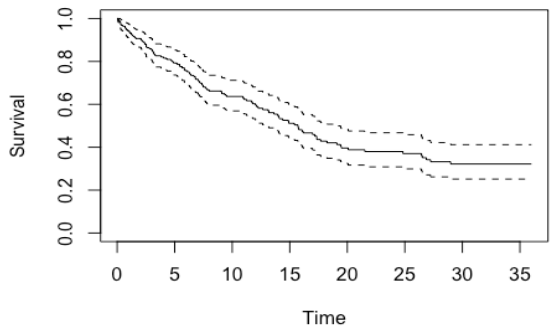

Figure 4. Survival function and its confidence intervals

b) Compute and graph separate survival curves for those with and without a record of crime against persons; test for differences between the two survival functions.

In this case, I apply the function 'survfit' but only with 'personal' covariate in order to see the effect of this type of crime.

After that, I execute the low-rank test and check if there is a relevant difference between both groups of the 'personal' covariate.

```
fit2<-survfit(Surv(months, censor)~personal,data=hen)
summary(fit2)

plot(fit2,lty=1:2,col=1:2,xlab="Time",ylab="Survival")
legend('topright',legend=c('no persons','persons'),cex=0.8,lty=1:2, col=1:2)
```

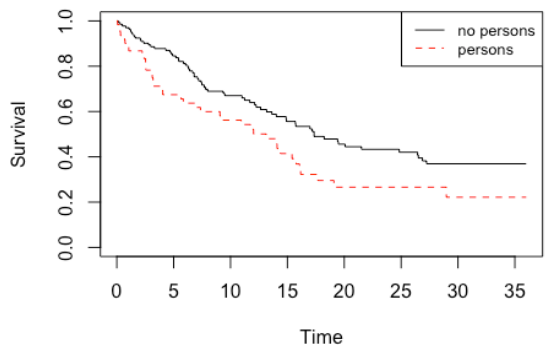

Figure 5. Survival function for personal crimes

```
#Low-rank test
survdiff(Surv(months, censor)~personal,data=hen)

## Call:
## survdiff(formula = Surv(months, censor) ~ personal, data = hen)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## personal=no   133       67     77.8      1.50       5.7
## personal=yes   61       39     28.2      4.14       5.7
##
##  Chisq= 5.7  on 1 degrees of freedom, p= 0.02
```

The p-value is 0.02 since it is lower than 0.05 (usually agreed alpha risk) we can reject the null hyphotesis which is "no effect on separated groups". Therefore, we can accept there is significance in treating both groups separately.

c) Compute and graph separate survival curves for those with and without a record of crime against property; test for differences between the two survival functions.

Here, I do the same if the last section but checking the relevance of the 'property' type of crime.

```
fit3<-survfit(Surv(months, censor)~property,data=hen)
summary(fit3)

plot(fit3,lty=1:2,col=1:2,xlab="Time",ylab="Survival")
legend('topright',legend=c('no property','property'),cex=0.5,lty=1:2, col=1:2)
```

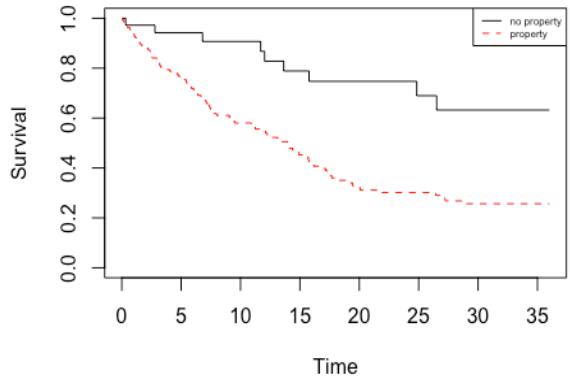

Figure 5. Survival function for property crimes

```
#Low-rank test
survdiff(Surv(months, censor)~property,data=hen)

## Call:
## survdiff(formula = Surv(months, censor) ~ property, data = hen)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## property=no    36        9     24.7      9.97      13.1
## property=yes 158       97     81.3      3.02      13.1
##
##  Chisq= 13.1  on 1 degrees of freedom, p= 3e-04
```

In this case, the p-value is 0.0003 since it is lower than 0.05 (usually agreed alpha risk) we can reject the null hyphotesis which is "no effect on separated groups". Therefore, we can accept there is significance in treating both groups separately.

Extra section)

Here, I apply the same than before but for the four different groups:

- Persons and Property

- Only Persons
- Only Property
- Other crime

```
fit4<-survfit(Surv(months, censor)~V7,data=hen)
summary(fit4)

plot(fit4,lty=1:4,col=1:4,xlab="Time",ylab="Survival")
legend('topright',legend=c('persons and property', 'persons','property', 'Other
crime'),cex=0.5,lty=1:4, col=1:4)
```

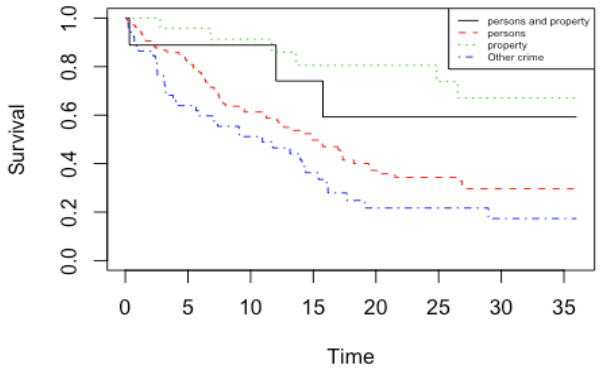

Figure 6. Survival function for the different crimes

```
#Low-rank test
survdiff(Surv(months, censor)~V7,data=hen)

## Call:
## survdiff(formula = Surv(months, censor) ~ V7, data = hen)
##
##                            N Observed Expected (O-E)^2/E (O-E)^2/V
## V7=Only Persons            9        3     5.36     1.036     1.095
## V7=Only Property         106       61    58.48     0.109     0.243
## V7=Other crime            27        6    19.33     9.191    11.387
## V7=Persons and Property   52       36    22.84     7.590     9.762
##
##  Chisq= 18.2  on 3 degrees of freedom, p= 4e-04
```

The p-value is 0.0004 since it is lower than 0.05 (usually agreed alpha risk) we can reject the null hyphotesis which is "no effect on separated groups". Therefore, we can accept there is significance in treating groups separately, at least, two groups.

d) Fit a Cox regression of time to re-arrest on the covariates personal, property, and cage.

(1) Determine by a Wald test whether each estimated coefficient is statistically significant.

In order to obtain the cox regression, I use the function 'coxph' and then we check the summary of the fit to know the value of the wald test and the significance of the covariates.

```
fit.all=coxph(Surv(months,censor)~ personal + property + cage, hen)
fit.all

## Call:
## coxph(formula = Surv(months, censor) ~ personal + property +
##     cage, data = hen)
##
##                 coef exp(coef) se(coef)      z        p
## personalyes  0.56914   1.76674  0.20521  2.773  0.00555
## propertyyes  0.93579   2.54922  0.35088  2.667  0.00765
## cage        -0.06671   0.93546  0.01678 -3.976 7.01e-05
##
## Likelihood ratio test=38.96  on 3 df, p=1.77e-08
## n= 194, number of events= 106

summary(fit.all)

## Call:
## coxph(formula = Surv(months, censor) ~ personal + property +
##     cage, data = hen)
##
##   n= 194, number of events= 106
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## personalyes  0.56914   1.76674  0.20521  2.773  0.00555 **
## propertyyes  0.93579   2.54922  0.35088  2.667  0.00765 **
## cage        -0.06671   0.93546  0.01678 -3.976 7.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## personalyes    1.7667     0.5660    1.1817    2.6415
## propertyyes    2.5492     0.3923    1.2816    5.0708
## cage           0.9355     1.0690    0.9052    0.9667
##
## Concordance= 0.694  (se = 0.027 )
## Likelihood ratio test= 38.96  on 3 df,   p=2e-08
## Wald test            = 29.02  on 3 df,   p=2e-06
## Score (logrank) test = 30.3  on 3 df,   p=1e-06
```

The wald test says that the p-value is 0.000002, so the model is significant and it has significant covariates. To see which are more significant, we check the p-value of the $Pr(>|z|)$ column for each predictor, and we can see, all of them are highly significant because they are much lower than 0.05.

(2) Interpret each of the estimated Cox-regression coefficient.

In order to interpret the regression, we have to look at the sign of the regression coefficients (coef) in the summary.

A positive sign means that the hazard (risk of being re-arrested) is higher, therefore, the prediction is worse for subjects with higher values of that variable.

The variable personal is encoded as a factor, no: didn't committed personal crime, yes: committed personal crime.

The R summary for the Cox model gives the coef for the second group relative to the first group, that is, personal crime versus no personal crime. The beta coefficient for personal = 0.57 indicates that who committed personal crime have higher risk of being re-arrested.

Regarding the property variable, it works as the same way than the personal variable, then, in this case the coef is equal to 0.94, this means that to have a crime against a property also increase the recidivism possibility because this value is positve.

5. Given a hazard function h(t) = c, where c > 0, derive the survival and the density function. Calculate the median failure time for c = 5.

- Survival function (s(t)):

$$H(t) = \int_0^t h\,(t)\;\partial t = \int_0^t c\;\;\partial t = c \cdot t\Big]_0^t = c\;t$$

$$S(t) = \exp(-c\;t)$$

- Density function (f(t)):

$$h(t) = \frac{f(t)}{s(t)} \Rightarrow f(t) = h(t) \cdot s(t)$$

$$f(t) = c \cdot \exp(-c\;t)$$

If c = 5, then,

$$f(t) = 5 \cdot \exp(-5t)$$

$$\text{median} = \frac{\ln(2)}{\lambda} = \frac{\ln(2)}{5} = 0.139$$

6. Consider the lung cancer data available from https://www.mayo.edu/research/documents/lunghtml/DOC-10027247.

The data set contains the following variables: - Enrolling institution Survival time - Status 1=alive, 2=dead Age - Sex 1=male 2=female - ECOG performace score, as judged by

physician: 0,1,2,3 - Karnofsky performace score, as judged by physician: 100, 90, . . ., 30 Karnofsky performace score, as judged by the patient (self) - Daily calories consumed at meals - Weight loss in the last 30 days (negative number = weight gain)

With this lung cancer data set, perform the analysis by following the steps listed below:

a) Fit a Cox PH model with all covariates included. For ECOG, you may simply treat it as continuous. For missing values, apply listwise deletion.

Firstly, I load the data, complete the header, delete the missing values and change the variable types to have them as numeric instead of as factor.

```
cancer <- read.table('cancer.txt', header=F)
colnames(cancer) <- c("inst","time","status","age","sex","ph_ecog","phy_karno","
pat_karno","meal_cal","wt_loss")

#Selecting the missing values and deleting them by listwise deletion
miss <- sum(cancer=='.')

cancer[cancer=='.']='NA'

miss2 <- sum(is.na(cancer))
cancer <- na.omit(cancer)

#Definig and changing variable types
cancer[,c(1:4,6:10)] <- sapply(cancer[,c(1:4,6:10)],as.character)
cancer[,c(1:4,6:10)] <- sapply(cancer[,c(1:4,6:10)],as.numeric)
```

We have also changed first the variable 'status' to identify the ones that didn't die as censored data (0) and the people who die as not censored (1).

then, we apply the fucntion 'coxph' to obtain the cox regression model.

```
fit.coxph=coxph(Surv(time, status) ~ ., cancer)
fit.coxph

## Call:
## coxph(formula = Surv(time, status) ~ ., data = cancer)
##
##                 coef  exp(coef)   se(coef)      z        p
## inst      -3.037e-02  9.701e-01  1.312e-02 -2.315 0.020619
## age        1.281e-02  1.013e+00  1.194e-02  1.073 0.283403
## sex       -5.666e-01  5.674e-01  2.014e-01 -2.814 0.004890
## ph_ecog    9.074e-01  2.478e+00  2.386e-01  3.803 0.000143
## phy_karno  2.658e-02  1.027e+00  1.163e-02  2.286 0.022231
## pat_karno -1.091e-02  9.891e-01  8.141e-03 -1.340 0.180160
## meal_cal   2.602e-06  1.000e+00  2.677e-04  0.010 0.992244
## wt_loss   -1.671e-02  9.834e-01  7.911e-03 -2.112 0.034647
##
## Likelihood ratio test=33.7  on 8 df, p=4.609e-05
## n= 167, number of events= 120
```

```
summary(fit.coxph)

## Call:
## coxph(formula = Surv(time, status) ~ ., data = cancer)
##
##   n= 167, number of events= 120
##
##                 coef  exp(coef)   se(coef)      z Pr(>|z|)
## inst      -3.037e-02  9.701e-01  1.312e-02 -2.315 0.020619 *
## age        1.281e-02  1.013e+00  1.194e-02  1.073 0.283403
## sex       -5.666e-01  5.674e-01  2.014e-01 -2.814 0.004890 **
## ph_ecog    9.074e-01  2.478e+00  2.386e-01  3.803 0.000143 ***
## phy_karno  2.658e-02  1.027e+00  1.163e-02  2.286 0.022231 *
## pat_karno -1.091e-02  9.891e-01  8.141e-03 -1.340 0.180160
## meal_cal   2.602e-06  1.000e+00  2.677e-04  0.010 0.992244
## wt_loss   -1.671e-02  9.834e-01  7.911e-03 -2.112 0.034647 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## inst         0.9701     1.0308    0.9455    0.9954
## age          1.0129     0.9873    0.9895    1.0369
## sex          0.5674     1.7623    0.3824    0.8420
## ph_ecog      2.4778     0.4036    1.5523    3.9552
## phy_karno    1.0269     0.9738    1.0038    1.0506
## pat_karno    0.9891     1.0110    0.9735    1.0051
## meal_cal     1.0000     1.0000    0.9995    1.0005
## wt_loss      0.9834     1.0169    0.9683    0.9988
##
## Concordance= 0.648  (se = 0.03 )
## Likelihood ratio test= 33.7  on 8 df,   p=5e-05
## Wald test            = 31.72  on 8 df,   p=1e-04
## Score (logrank) test = 32.51  on 8 df,   p=8e-05

ggforest(fit.coxph,cancer)
```

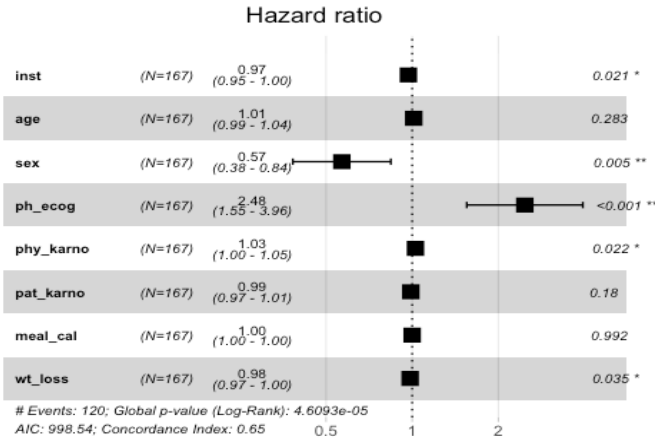

Figure 7. Forest graph of the significance of the variables and its confidence intervals

The p-value of the general model is 0.00005, while the values of the Likelihood ratio test is 33.70 and Wald test test, 31.72.

In this model, the p-value which describes the global significance of the model is lower than 0.05, then, we can reject the null hyphotesis (the coefficients are equal to 0 and they don't have effect in the regression), and accept the alternative one which is the coefficients are different than 0 and therefore, the model is significant.

b) Concerning the full Cox PH model,test to see if we can drop Karnofsky.physician and Karnofsky.patient simultaneously by using Wald test and likelihood ra-tio test (LRT). Compare the two testing results and comment.

Now, I do the same than the last section but I don't select the variables of Karnofsky as predictors.

```
fit.coxph2=coxph(Surv(time, status) ~ inst+age+sex+ph_ecog+meal_cal+wt_loss, can
cer)
fit.coxph2
```

```
## Call:
## coxph(formula = Surv(time, status) ~ inst + age + sex + ph_ecog +
##     meal_cal + wt_loss, data = cancer)
##
##                coef  exp(coef)   se(coef)      z        p
## inst     -2.736e-02  9.730e-01  1.303e-02 -2.100  0.03575
## age       6.742e-03  1.007e+00  1.140e-02  0.591  0.55422
## sex      -5.512e-01  5.763e-01  2.006e-01 -2.747  0.00601
## ph_ecog   6.218e-01  1.862e+00  1.544e-01  4.027 5.66e-05
## meal_cal -5.576e-05  9.999e-01  2.584e-04 -0.216  0.82916
## wt_loss  -1.403e-02  9.861e-01  7.751e-03 -1.809  0.07037
##
## Likelihood ratio test=27.01  on 6 df, p=0.0001444
## n= 167, number of events= 120
```

```
summary(fit.coxph2)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ inst + age + sex + ph_ecog +
##     meal_cal + wt_loss, data = cancer)
##
##   n= 167, number of events= 120
##
##                coef  exp(coef)   se(coef)      z Pr(>|z|)
## inst     -2.736e-02  9.730e-01  1.303e-02 -2.100  0.03575 *
## age       6.742e-03  1.007e+00  1.140e-02  0.591  0.55422
## sex      -5.512e-01  5.763e-01  2.006e-01 -2.747  0.00601 **
## ph_ecog   6.218e-01  1.862e+00  1.544e-01  4.027 5.66e-05 ***
## meal_cal -5.576e-05  9.999e-01  2.584e-04 -0.216  0.82916
## wt_loss  -1.403e-02  9.861e-01  7.751e-03 -1.809  0.07037 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##          exp(coef) exp(-coef) lower .95 upper .95
## inst        0.9730     1.0277    0.9485    0.9982
## age         1.0068     0.9933    0.9845    1.0295
## sex         0.5763     1.7353    0.3889    0.8539
## ph_ecog     1.8623     0.5370    1.3759    2.5206
## meal_cal    0.9999     1.0001    0.9994    1.0005
## wt_loss     0.9861     1.0141    0.9712    1.0012
##
## Concordance= 0.643  (se = 0.031 )
## Likelihood ratio test= 27.01  on 6 df,   p=1e-04
## Wald test            = 25.79  on 6 df,   p=2e-04
## Score (logrank) test = 26.27  on 6 df,   p=2e-04
```

The p-value of the general model is 0.0001, while the values of the Likelihood ratio test is 27.01 and Wald test test, 25.79.

In this second model, the p-value still lower than 0.05, so the model remains significant. Also, the values for the LRT and Wald test are smaller, so it means the model is better fitted. We can conclude, then, both dismissed variables are not important at all and can be removed simultaneously.

Regarding the comparison of LRT and Wald test, we can say they are asymptotically equivalent, so for large N, they will give similar results, but for small N, they may differ. In any case, LRT usually has better performance with small samples and then it is prefered in this case.

c) Find the best Cox model using a variable selection procedure of your choice.

As we can see in the summary and in the forest graph (figure 8) of the whole model, the clearly significant variables are sex and ph_ecog, however, let's evaluate further the rest of the variables checking their confidence intervals:

- inst: 0.95-1.00
- age: 0.99-1.04
- phy_karno: 1.00-1.05
- pat_karno: 0.97-1.01
- wt_loss: 0.97-1.00

Since all of them include the 1 inside the interval, we can assume none of them is significant.

Then, the best model to fit is the one that include the predictors: sex and ph_ecog.

```
fit.coxph3=coxph(Surv(time, status) ~ sex+ph_ecog, cancer)
fit.coxph3

## Call:
## coxph(formula = Surv(time, status) ~ sex + ph_ecog, data = cancer)
##
##             coef exp(coef) se(coef)      z        p
```

```
## sex     -0.5101    0.6004   0.1969 -2.591 0.009579
## ph_ecog  0.4825    1.6201   0.1323  3.647 0.000266
##
## Likelihood ratio test=19.48  on 2 df, p=5.882e-05
## n= 167, number of events= 120

summary(fit.coxph3)

## Call:
## coxph(formula = Surv(time, status) ~ sex + ph_ecog, data = cancer)
##
##   n= 167, number of events= 120
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## sex     -0.5101    0.6004   0.1969 -2.591 0.009579 **
## ph_ecog  0.4825    1.6201   0.1323  3.647 0.000266 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## sex        0.6004     1.6655    0.4082    0.8832
## ph_ecog    1.6201     0.6172    1.2501    2.0998
##
## Concordance= 0.641  (se = 0.031 )
## Likelihood ratio test= 19.48  on 2 df,   p=6e-05
## Wald test            = 19.35  on 2 df,   p=6e-05
## Score (logrank) test = 19.62  on 2 df,   p=5e-05
```

d) Interpret the final model in terms of hazard ratios. In particular, the 95% confidence intervals should be supplied for the hazard ratios.

Regarding the interpretation of the final model, we can go to the hazard ratios. They are the exponentiated coefficients (exp(coef)) of each variable.

The HR for sex2 is 0.6 (exp(-0.51)), since it is categorical, that value refers to the female sex and it gives the effect size of the covariate. This value reduces the hazard by a factor of 0.6 or, in other words, by 40%. Then, being female has lower risk of death.

On the other hand, the HR for ph_ecog is 1.62, since it is continuos, that value refers to effect of one more unit in that covariate. This value increase the hazard by a factor of 1.62, in other words, by 62%. Then, to have more performace score increase the risk of death.

e) Suppose that we want to test at significance level $\alpha = 0{,}05$ if the survival functions are the same between males and females, while adjusting for ECOG performance score. Since ECOG is measured on an ordinal scale, the adjust- ment could be made in three ways by

- *Treating ECOG as continuous.*
- *Treating ECOG as categorical by coding it with dummy variables*

- *Using the stratified Cox PH model approach.*

Carry out all three approaches and inspect how different the results are in terms of the (adjusted) gender effect.

We already have the model with the variable ph_ecog as continuous, then, let's see the result with the variable as categorical:

```
cancer[,6] <- as.factor(cancer[,6])

fit.coxph4=coxph(Surv(time, status) ~ sex+ph_ecog, cancer)
fit.coxph4

## Call:
## coxph(formula = Surv(time, status) ~ sex + ph_ecog, data = cancer)
##
##             coef exp(coef) se(coef)      z        p
## sex      -0.4998    0.6067   0.1974 -2.532 0.011353
## ph_ecog1  0.3206    1.3780   0.2331  1.375 0.168983
## ph_ecog2  0.9185    2.5056   0.2608  3.521 0.000429
## ph_ecog3  1.9973    7.3694   1.0357  1.929 0.053789
##
## Likelihood ratio test=20.39  on 4 df, p=0.000418
## n= 167, number of events= 120

summary(fit.coxph4)

## Call:
## coxph(formula = Surv(time, status) ~ sex + ph_ecog, data = cancer)
##
##   n= 167, number of events= 120
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## sex      -0.4998    0.6067   0.1974 -2.532 0.011353 *
## ph_ecog1  0.3206    1.3780   0.2331  1.375 0.168983
## ph_ecog2  0.9185    2.5056   0.2608  3.521 0.000429 ***
## ph_ecog3  1.9973    7.3694   1.0357  1.929 0.053789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## sex         0.6067     1.6483    0.4120    0.8933
## ph_ecog1    1.3780     0.7257    0.8726    2.1759
## ph_ecog2    2.5056     0.3991    1.5028    4.1777
## ph_ecog3    7.3694     0.1357    0.9680   56.1056
##
## Concordance= 0.646  (se = 0.03 )
## Likelihood ratio test= 20.39  on 4 df,   p=4e-04
## Wald test            = 21.86  on 4 df,   p=2e-04
## Score (logrank) test = 23.52  on 4 df,   p=1e-04
```

The effect of treating the variable 'ph_ecog' as categorical in the gender is almost null and the interpretation of the gender covariate remains the same.

On the other hand, let's see how the stratification affect:

```
fit.coxph5=coxph(Surv(time, status) ~ sex+strata(ph_ecog), cancer)
fit.coxph5

## Call:
## coxph(formula = Surv(time, status) ~ sex + strata(ph_ecog), data = cancer)
##
##        coef exp(coef) se(coef)      z       p
## sex -0.5434    0.5807   0.2002 -2.715 0.00664
##
## Likelihood ratio test=7.75  on 1 df, p=0.00537
## n= 167, number of events= 120

summary(fit.coxph5)

## Call:
## coxph(formula = Surv(time, status) ~ sex + strata(ph_ecog), data = cancer)
##
##   n= 167, number of events= 120
##
##        coef exp(coef) se(coef)      z Pr(>|z|)
## sex -0.5434    0.5807   0.2002 -2.715  0.00664 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## sex    0.5807      1.722    0.3923    0.8598
##
## Concordance= 0.572  (se = 0.027 )
## Likelihood ratio test= 7.75  on 1 df,   p=0.005
## Wald test            = 7.37  on 1 df,   p=0.007
## Score (logrank) test = 7.54  on 1 df,   p=0.006
```

In this case, the effect of treat the variable 'ph_ecog' as strata in the gender is a litle bit relevant, and now the interpretation of the gender covariate is that its HR value is 0.58 and it means that the length of survival increase by a factor of 0.58 or, in other words by 42%.