QuantProject

```
## Loading required package: ggpubr
```

```
## Loading required package: magrittr
```

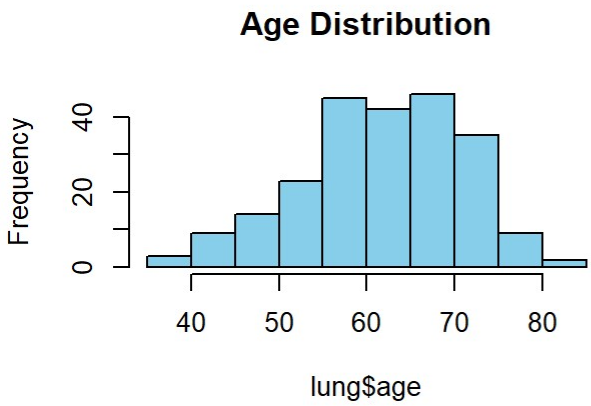
```
## Loading required package: lattice
```

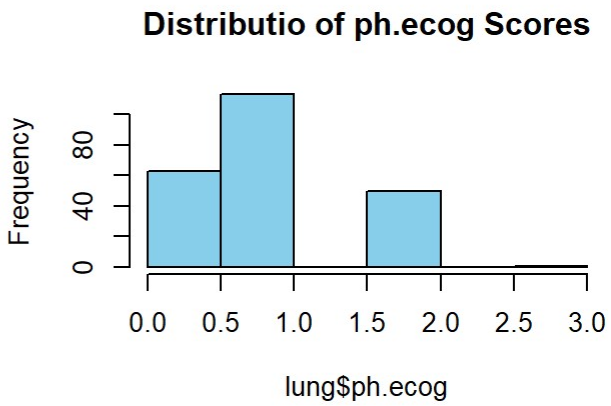
```
##
## Attaching package: 'caret'
```

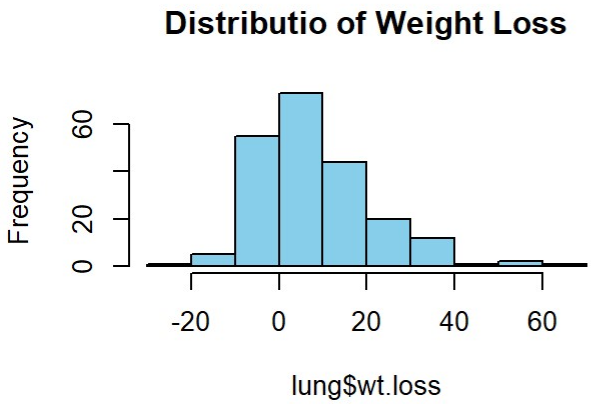
```
## The following object is masked from 'package:survival':
##
##     cluster
```

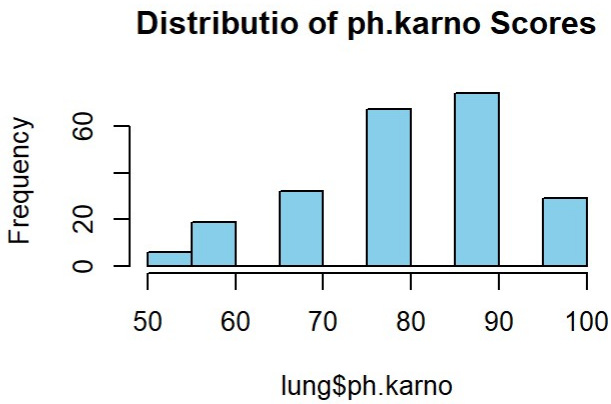
```
## [1] 2 2 1 2 2 1
```

```
par(mfrow = c(2,2))
hist(lung$age, main = "Age Distribution", col = "skyblue")
hist(lung$ph.ecog, main = "Distributio of ph.ecog Scores", col = "skyblue")
hist(lung$wt.loss, main = "Distributio of Weight Loss", col = "skyblue")
hist(lung$ph.karno, main = "Distributio of ph.karno Scores", col = "skyblue")
```

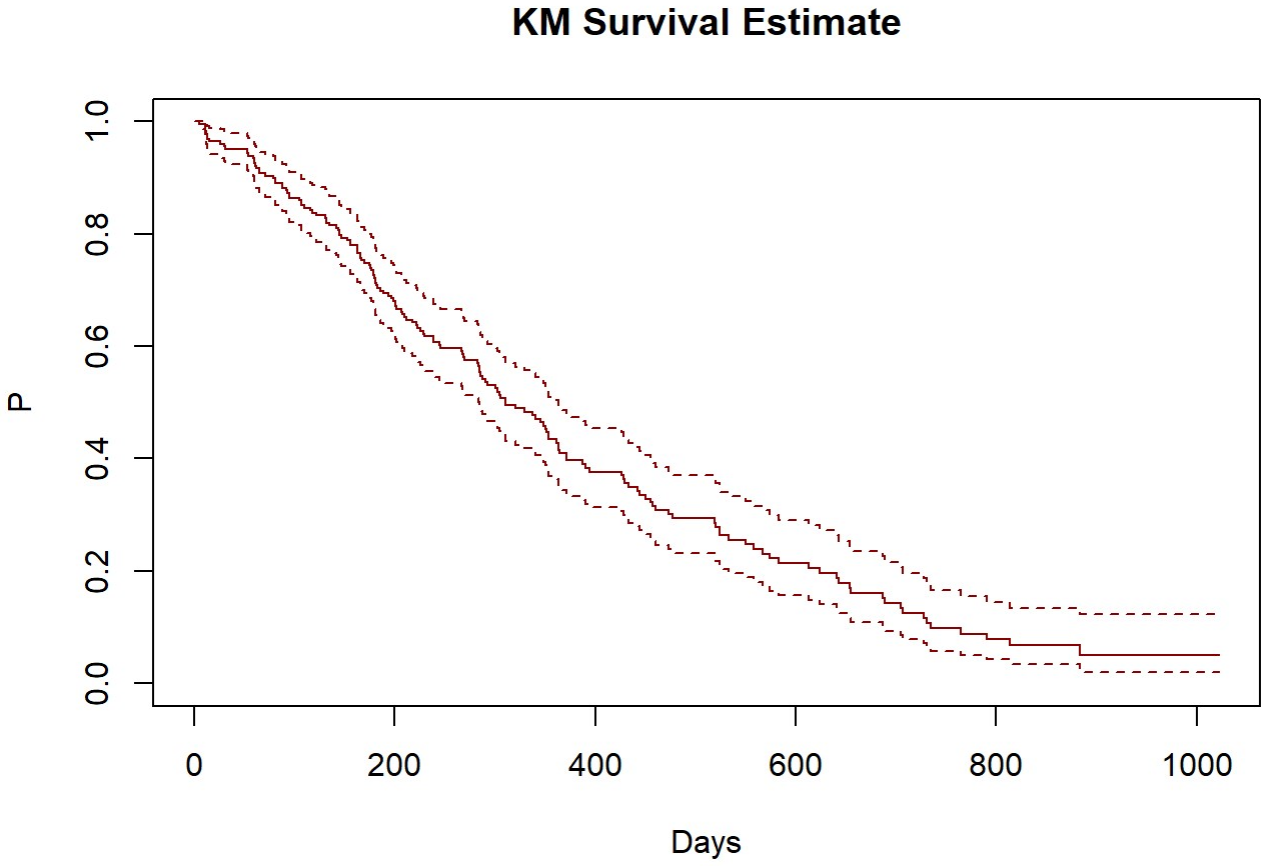




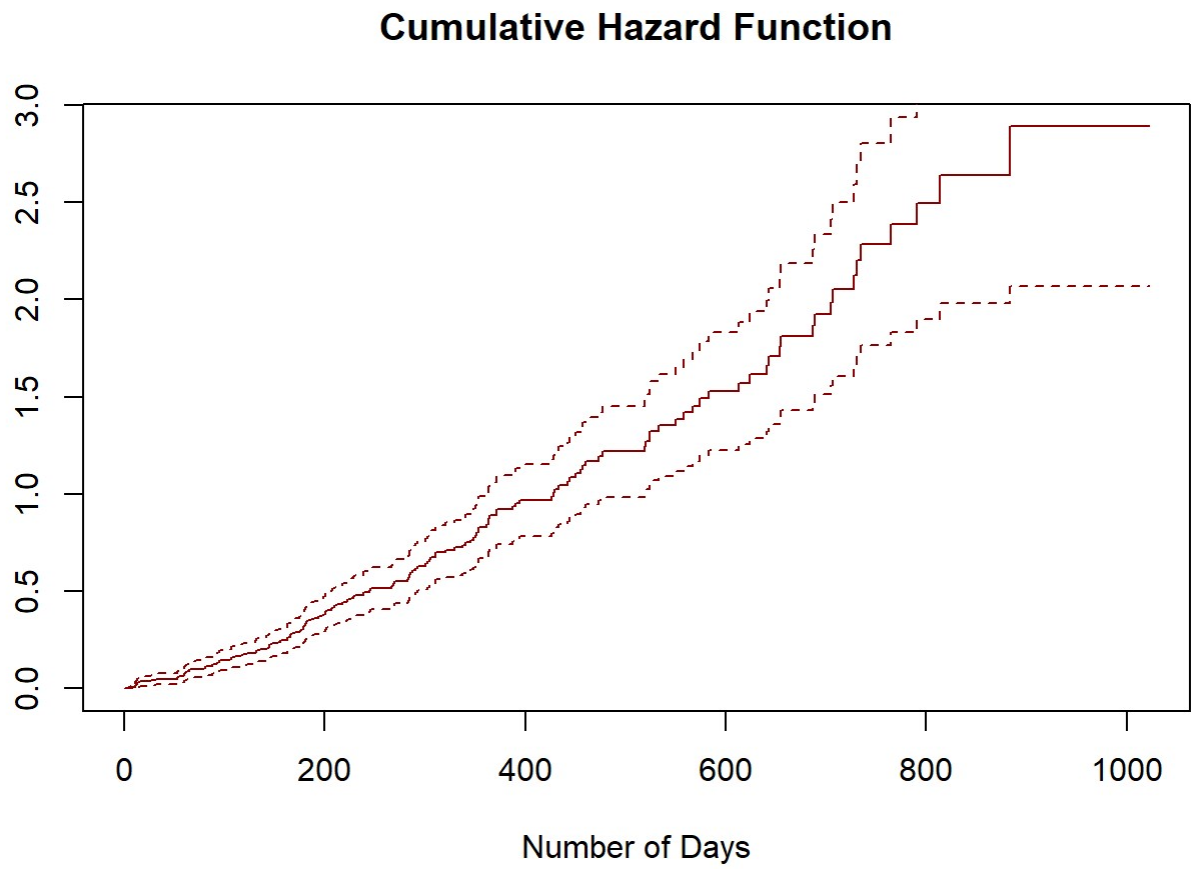




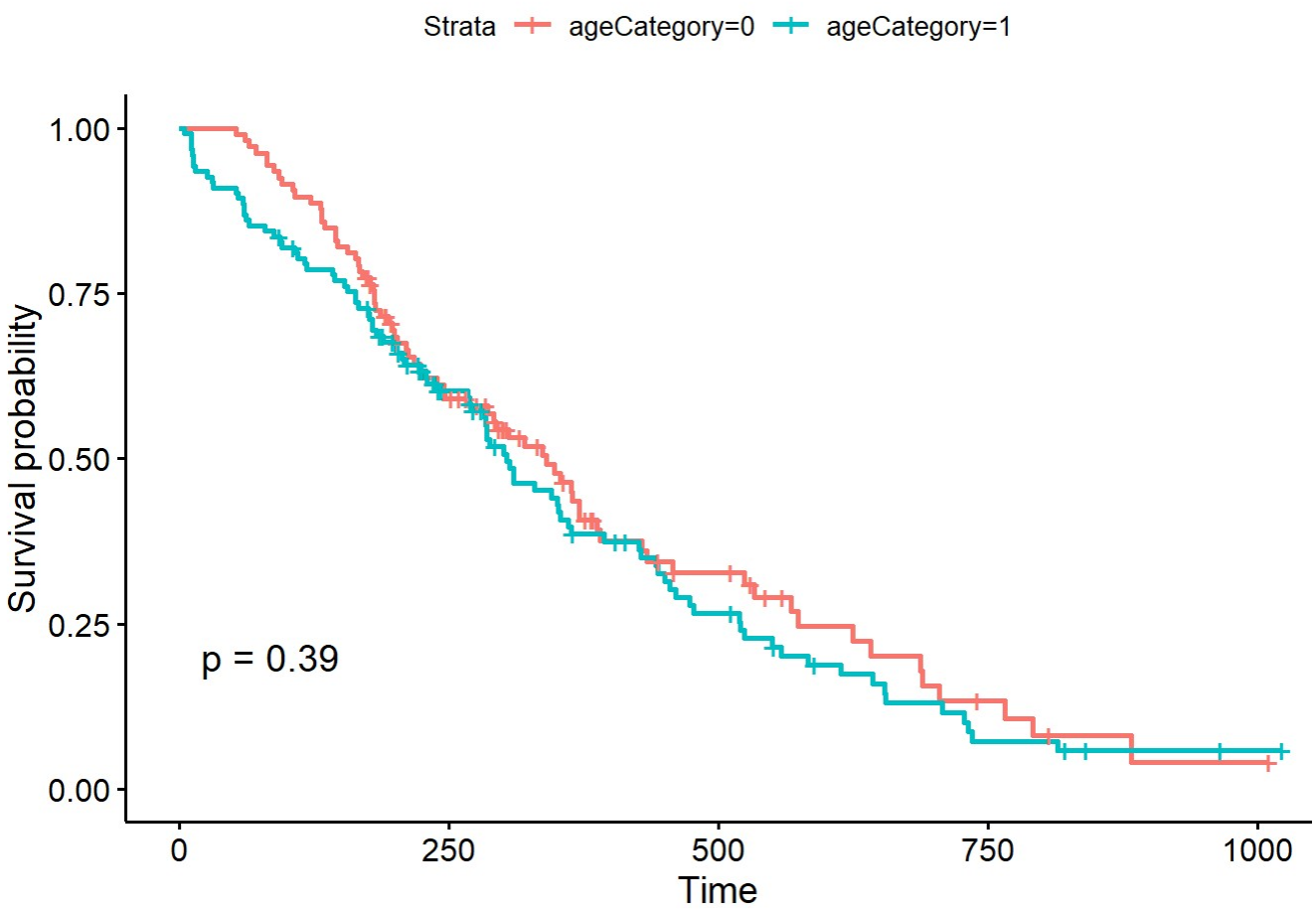

```
surv <- Surv(time = lung$time, event = lung$status)
all <- survfit(surv ~ 1, data = lung)
plot(all, main = "KM Survival Estimate", xlab = "Days", ylab = "P", col = "darkred")
```



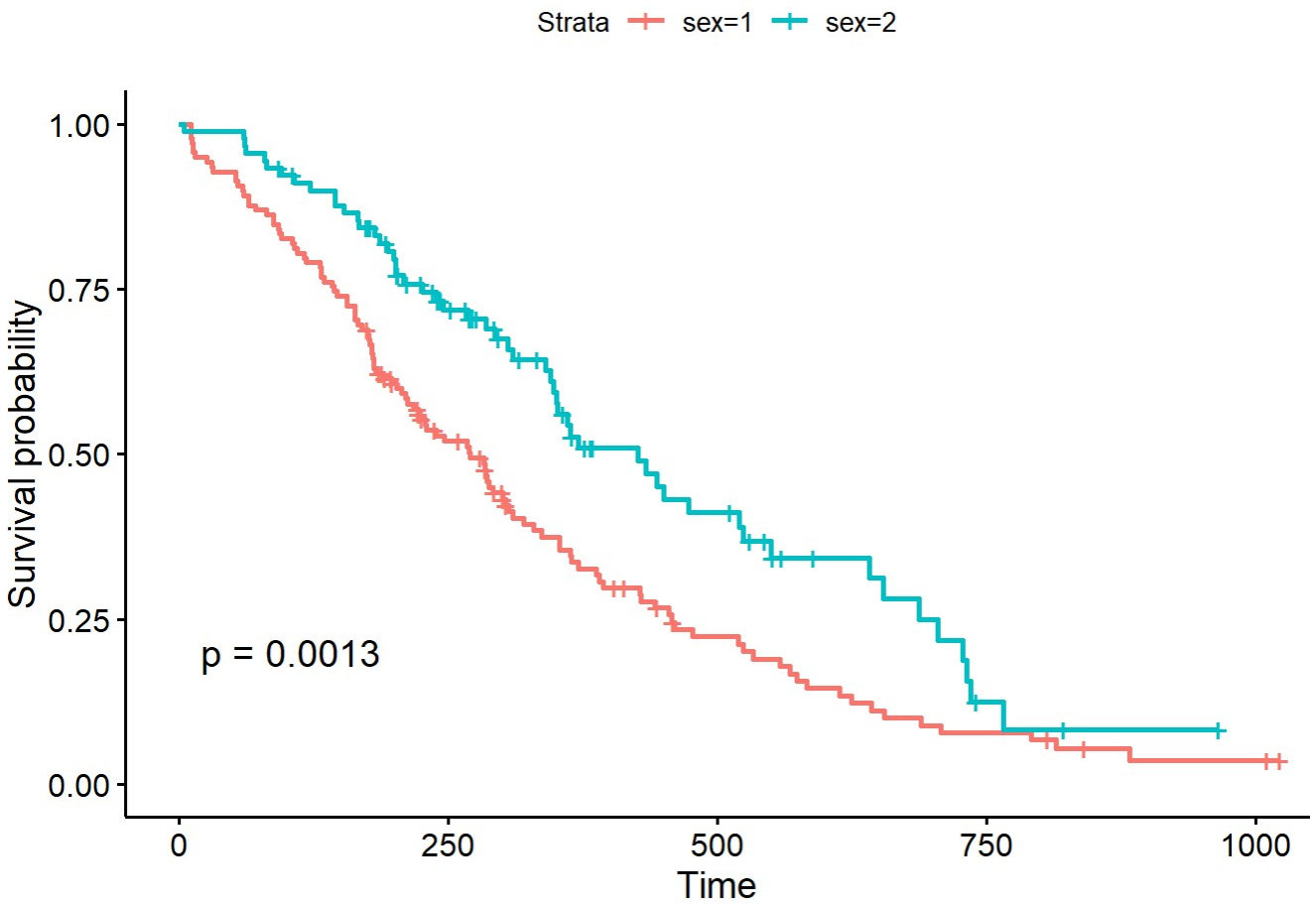

```
#cumulative hazard function
hazard <- plot(all, fun="cumhaz", col = "darkred", main = "Cumulative Hazard Function
", xlab = "Number of Days")
```



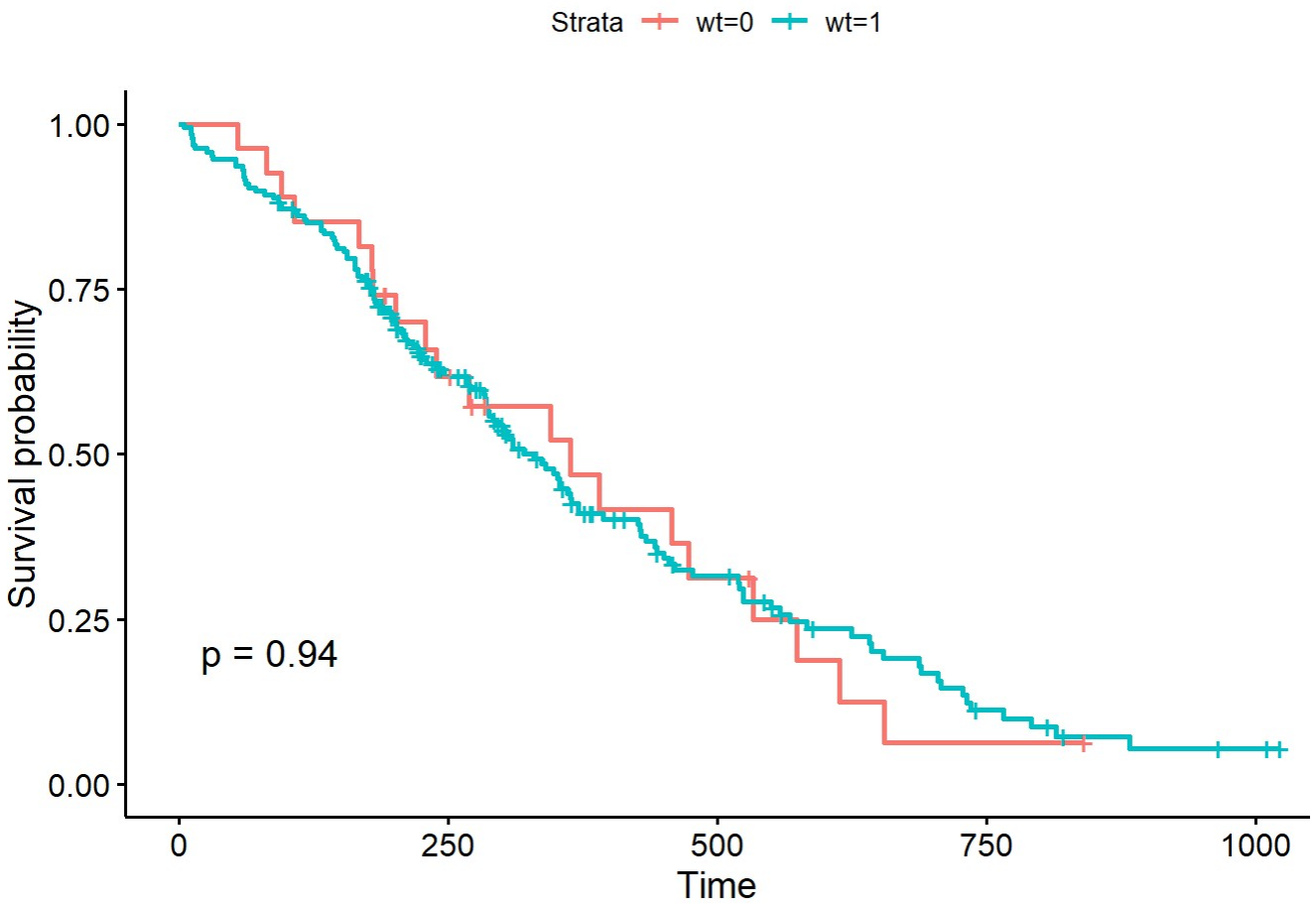

```
#surival function of age
#divided into two groups those over and under the mean age
ageCategory <- as.numeric(lung$age>mean(lung$age))
age <- survfit(surv ~ ageCategory, data = lung)
ageplot <-ggsurvplot(age,data = lung, pval = TRUE,linetype = "solid")
ageplot
```



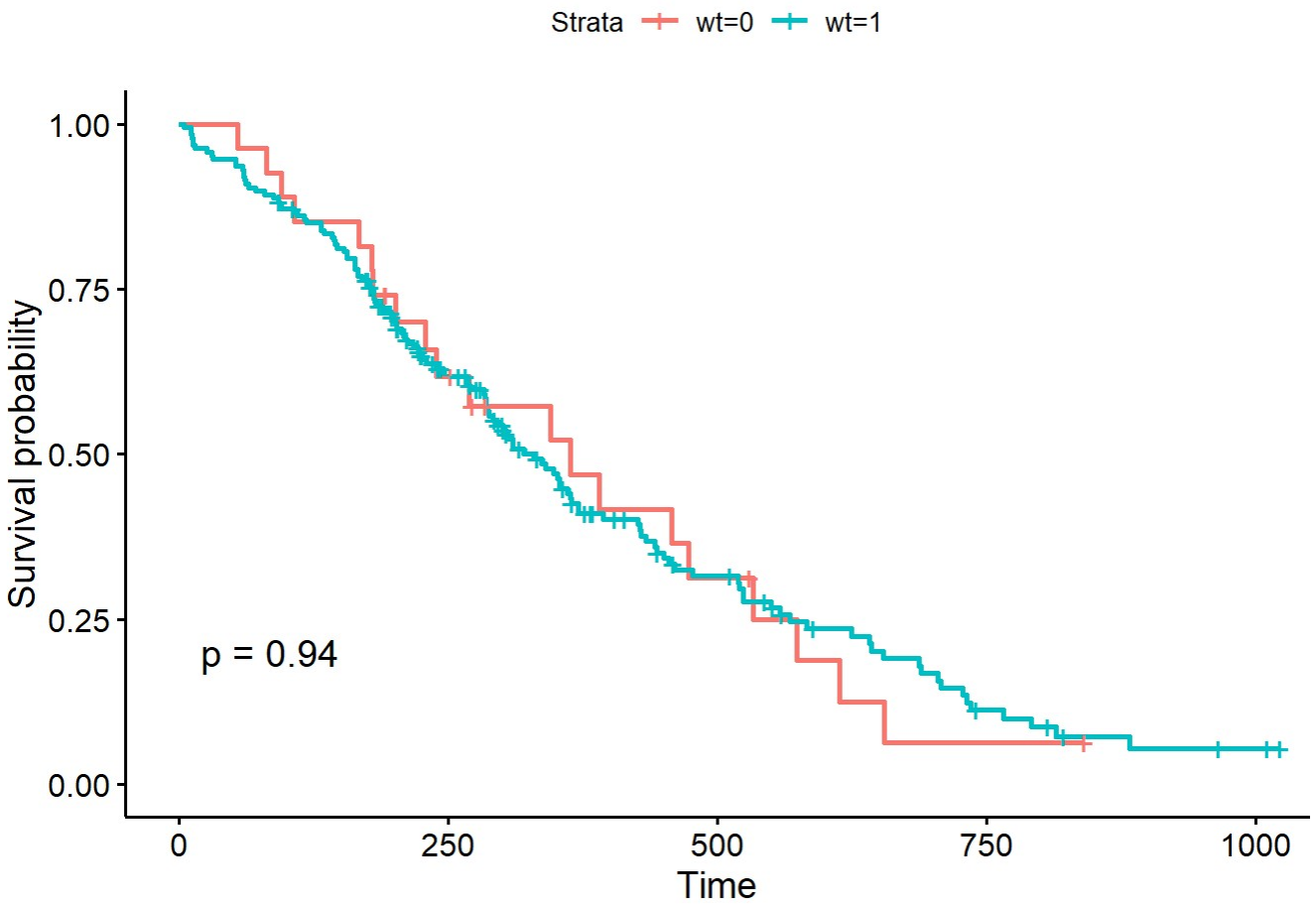

```
#survival plot for sex

sexFit <- survfit(surv ~ sex, data = lung)
sexplot <- ggsurvplot(sexFit,data = lung, pval = TRUE,linetype = "solid")
sexplot
```



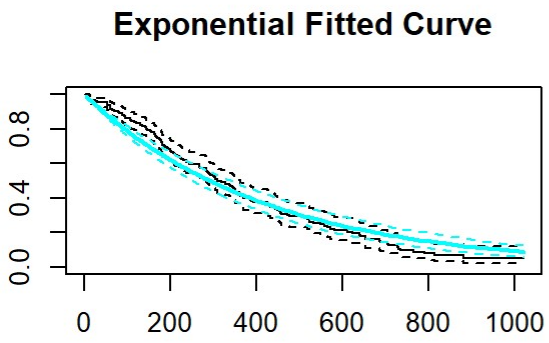

```
#weight loss, 1 = gained or no loss, 0 = lost weight
wt <- as.numeric(lung$wt.loss >= 0)
wtFit <- survfit(surv ~ wt, data = lung)
wtplot <- ggsurvplot(wtFit,data = lung, pval = TRUE,linetype = "solid")
wtplot
```

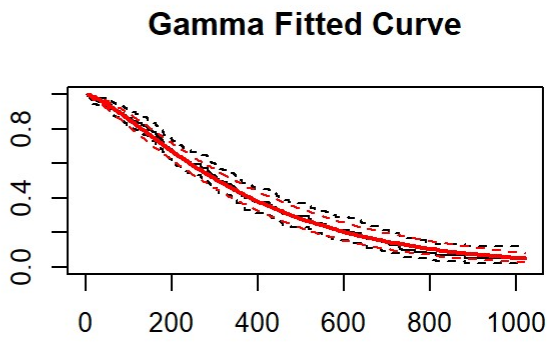


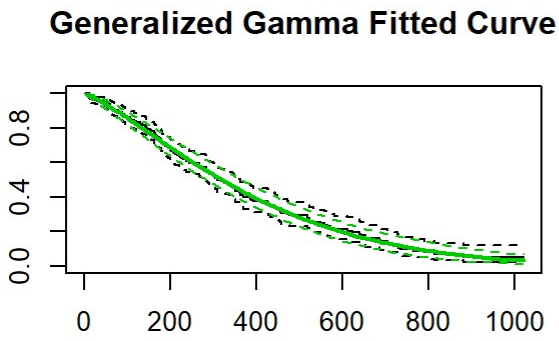

```
#weight loss, 1 = gained or no loss, 0 = lost weight
wt <- as.numeric(lung$wt.loss >= 0)
wtFit <- survfit(surv ~ wt, data = lung)
wtplot <- ggsurvplot(wtFit,data = lung, pval = TRUE,linetype = "solid")
wtplot
```

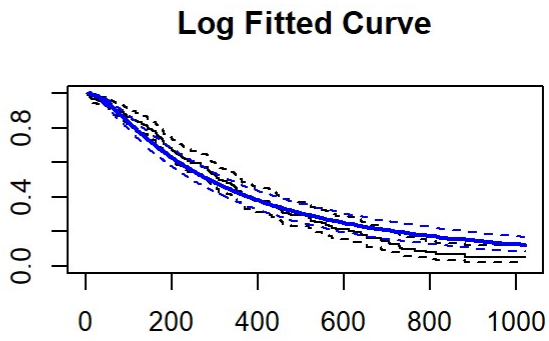



```
library(flexsurv)
exp <- flexsurvreg(surv~1, dist="exp")
gamma <- flexsurvreg(surv~1, dist="gamma")
genGamma <- flexsurvreg(surv~1, dist="gengamma")
log <- flexsurvreg(surv~1, dist="lognormal")

AICs <- c(exp$AIC, gamma$AIC, genGamma$AIC, log$AIC)
logs <- c(exp$loglik, gamma$loglik, genGamma$loglik, log$loglik)
dists <- c("Exponential", "Gamma", "Generalized Gammal", "log")
table <- rbind(dists, AICs, logs)


par(mfrow = c(2,2))
plot(exp,col = 5,main = "Exponential Fitted Curve")

plot(gamma, col = 2, main = "Gamma Fitted Curve")

plot(genGamma, col = 3, main = "Generalized Gamma Fitted Curve")

plot(log, col = 4, main = "Log Fitted Curve")
```










```
table
```

```
##       [,1]                [,2]                [,3]
## dists "Exponential"       "Gamma"             "Generalized Gammal"
## AICs  "2326.67635157493"  "2313.46926750336"  "2313.37959210286"
## logs  "-1162.33817578747" "-1154.73463375168" "-1153.68979605143"
##       [,4]
## dists "log"
## AICs  "2342.5381106112"
## logs  "-1169.2690553056"
```

Survival proportions can be accurately modeled using the gamma distrabution

```
#creating cox survival object
lung <-na.omit(lung)
survCox <- Surv(time = lung$time, event = lung$status)
```

```
#cox univariant model with sex only
coxSex <- coxph(survCox ~ lung$sex, data = lung)
summary(coxSex)
```

```
## Call:
## coxph(formula = survCox ~ lung$sex, data = lung)
##
##   n= 167, number of events= 120
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## lung$sex -0.4792    0.6193   0.1966 -2.437   0.0148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## lung$sex    0.6193      1.615    0.4212    0.9104
##
## Concordance= 0.567  (se = 0.025 )
## Likelihood ratio test= 6.25  on 1 df,   p=0.01
## Wald test            = 5.94  on 1 df,   p=0.01
## Score (logrank) test = 6.05  on 1 df,   p=0.01
```

```
#cox univariant model with age only
coxAge <-coxph(survCox ~ lung$age, data = lung)
summary(coxAge)
```

```
## Call:
## coxph(formula = survCox ~ lung$age, data = lung)
##
##   n= 167, number of events= 120
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## lung$age 0.01989   1.02009  0.01075 1.851   0.0642 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## lung$age      1.02     0.9803    0.9988     1.042
##
## Concordance= 0.559  (se = 0.03 )
## Likelihood ratio test= 3.52  on 1 df,   p=0.06
## Wald test            = 3.43  on 1 df,   p=0.06
## Score (logrank) test = 3.44  on 1 df,   p=0.06
```

```
#cox univariant model with age only
coxecog <-coxph(survCox ~ lung$ph.ecog, data = lung)
summary(coxecog)
```

```
## Call:
## coxph(formula = survCox ~ lung$ph.ecog, data = lung)
##
##   n= 167, number of events= 120
##
##                coef exp(coef) se(coef)     z Pr(>|z|)
## lung$ph.ecog 0.4693    1.5988   0.1331 3.527  0.00042 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## lung$ph.ecog     1.599     0.6255     1.232     2.075
##
## Concordance= 0.615  (se = 0.028 )
## Likelihood ratio test= 12.41  on 1 df,   p=4e-04
## Wald test            = 12.44  on 1 df,   p=4e-04
## Score (logrank) test = 12.61  on 1 df,   p=4e-04
```

```
#multivariate cox model with stepwise selection of variables
Train <- createDataPartition(lung$status, p=0.75, list=FALSE)
training <- lung[ Train, ]
testing <- lung[ -Train, ]


res.cox <- coxph(survCox ~ lung$age + lung$sex + lung$ph.ecog +lung$ph.karno + lung$p
at.karno + lung$meal.cal + lung$wt.loss, data =  training)
sp <- step(res.cox, data = training)
```

```
## Start:  AIC=1002.07
## survCox ~ lung$age + lung$sex + lung$ph.ecog + lung$ph.karno +
##     lung$pat.karno + lung$meal.cal + lung$wt.loss
##
##                   Df    AIC
## - lung$meal.cal    1 1000.1
## - lung$age         1 1001.0
## <none>               1002.1
## - lung$pat.karno   1 1002.3
## - lung$wt.loss     1 1003.6
## - lung$ph.karno    1 1004.3
## - lung$sex         1 1008.0
## - lung$ph.ecog     1 1011.1
##
## Step:  AIC=1000.08
## survCox ~ lung$age + lung$sex + lung$ph.ecog + lung$ph.karno +
##     lung$pat.karno + lung$wt.loss
##
##                   Df     AIC
## - lung$age         1  998.95
## <none>               1000.08
## - lung$pat.karno   1 1000.29
## - lung$wt.loss     1 1001.60
## - lung$ph.karno    1 1002.28
## - lung$sex         1 1006.29
## - lung$ph.ecog     1 1009.09
##
## Step:  AIC=998.95
## survCox ~ lung$sex + lung$ph.ecog + lung$ph.karno + lung$pat.karno +
##     lung$wt.loss
##
##                   Df     AIC
## <none>               998.95
## - lung$pat.karno   1  999.34
## - lung$ph.karno    1 1000.53
## - lung$wt.loss     1 1000.74
## - lung$sex         1 1005.25
## - lung$ph.ecog     1 1007.83
```

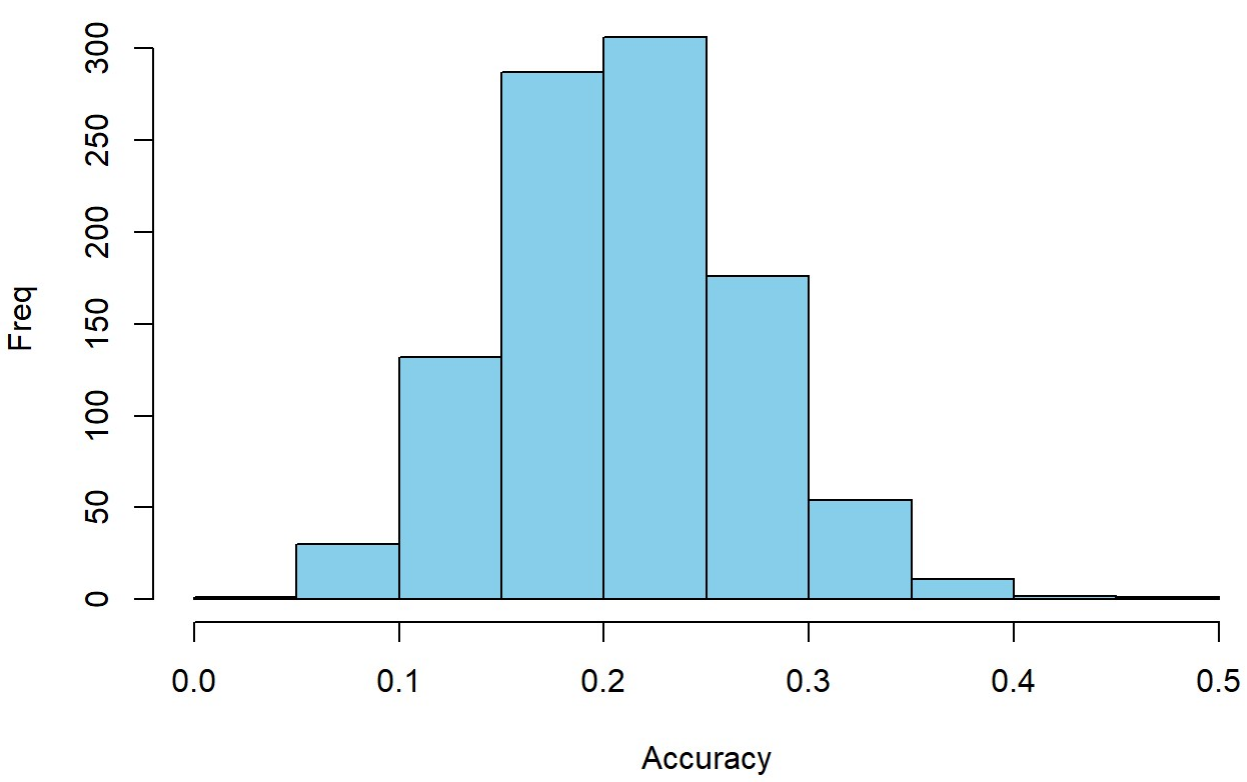
```
#updated model based on results of the step selection
update.cox <- coxph(survCox ~ lung$sex + lung$ph.ecog + lung$ph.karno +
    lung$pat.karno + lung$wt.loss, data = lung)
summary(update.cox)
```

```
## Call:
## coxph(formula = survCox ~ lung$sex + lung$ph.ecog + lung$ph.karno +
##     lung$pat.karno + lung$wt.loss, data = lung)
##
##   n= 167, number of events= 120
##
##                      coef exp(coef)  se(coef)      z Pr(>|z|)
## lung$sex       -0.558190  0.572244  0.199202 -2.802  0.00508 **
## lung$ph.ecog    0.742983  2.102197  0.227604  3.264  0.00110 **
## lung$ph.karno   0.020366  1.020575  0.011080  1.838  0.06604 .
## lung$pat.karno -0.012401  0.987675  0.007978 -1.554  0.12008
## lung$wt.loss   -0.014494  0.985611  0.007693 -1.884  0.05957 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## lung$sex          0.5722     1.7475    0.3873    0.8456
## lung$ph.ecog      2.1022     0.4757    1.3457    3.2841
## lung$ph.karno     1.0206     0.9798    0.9987    1.0430
## lung$pat.karno    0.9877     1.0125    0.9724    1.0032
## lung$wt.loss      0.9856     1.0146    0.9709    1.0006
##
## Concordance= 0.658  (se = 0.029 )
## Likelihood ratio test= 27.28  on 5 df,   p=5e-05
## Wald test            = 26.89  on 5 df,   p=6e-05
## Score (logrank) test = 27.64  on 5 df,   p=4e-05
```

```
k<-1000
acc <- NULL


for(i in 1:k){
Train <- createDataPartition(lung$status, p=0.75, list=FALSE)
training <- lung[ Train, ]
testing <- lung[ -Train, ]

testSurv <- Surv(time = training$time, event = training$status)

update.coxT <- coxph(testSurv ~ sex + ph.ecog + ph.karno +
   pat.karno + wt.loss, data = training)

pred<-predict(update.coxT, newdata=testing,type = "risk")


results <- ifelse(pred > 0.5,1,0)
answers <- testing$status
misClasificError <- mean(answers != results)
acc[i]=1-misClasificError
}
mean(acc)
```

```
## [1] 0.2162927
```

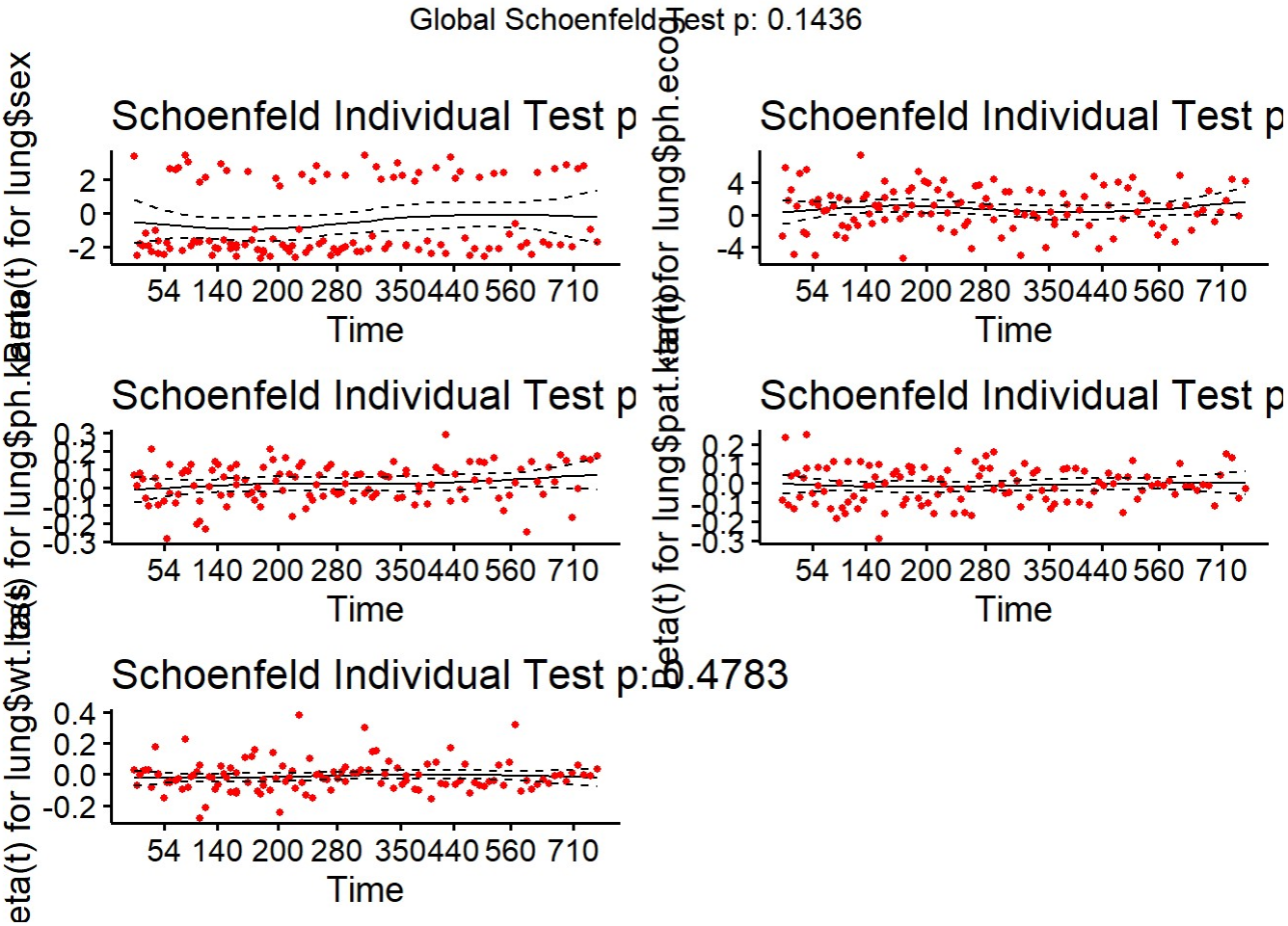
```
hist(acc,xlab='Accuracy',ylab='Freq',
     col='skyblue', main = "Accuracy of Cox Model 100 Iterations")
```




```
#testing cox model assumptions

assump <- cox.zph(update.cox)
print("From these tests we find that we can consider all of the covarientes significa
nt ")
```

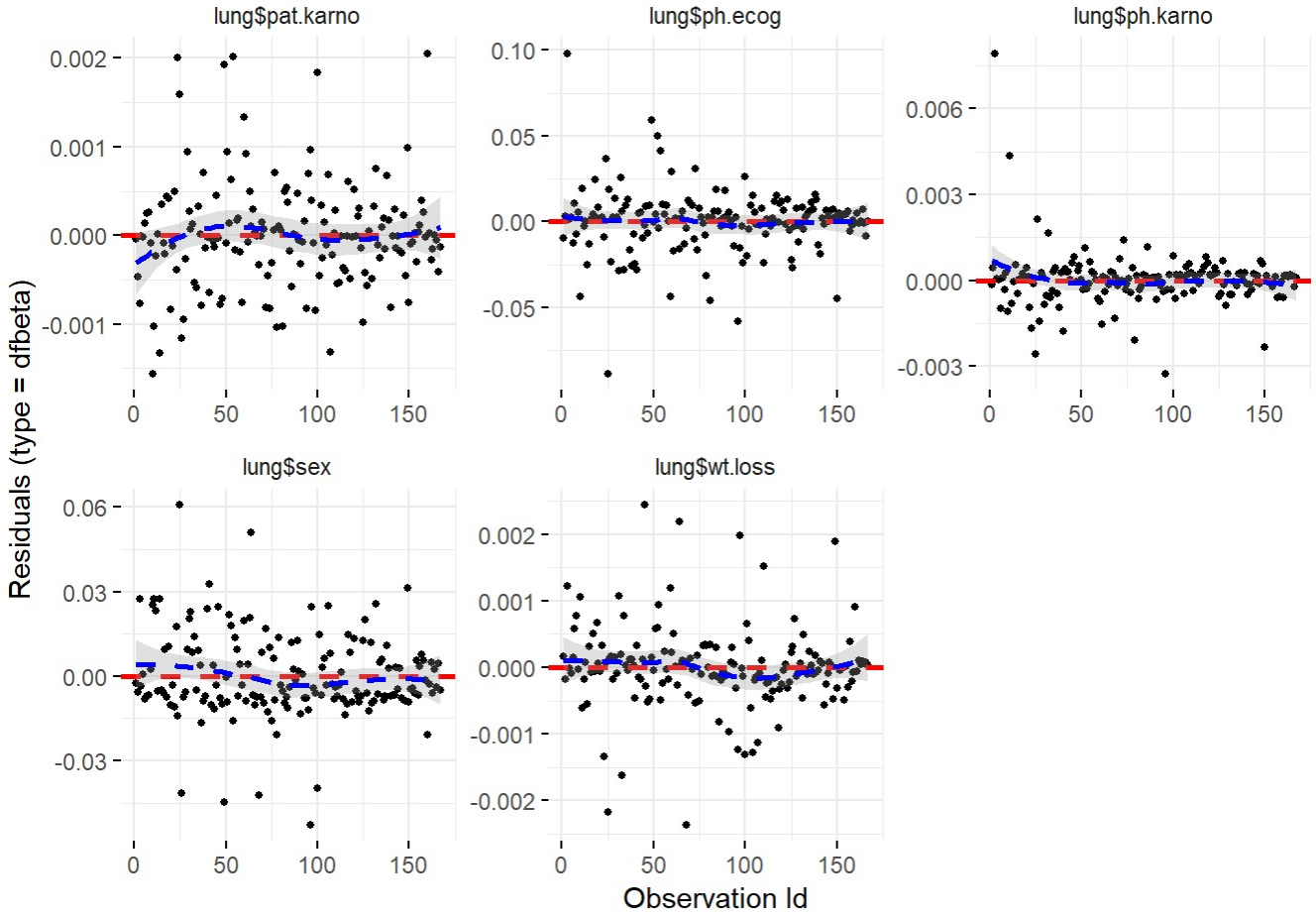
```
## [1] "From these tests we find that we can consider all of the covarientes signific
ant "
```

```
ggcoxzph(assump)
```




```
#plot of the Schoenfeld residuals
#alligns with assumptions and the results of the previous test
```

```
ggcoxdiagnostics(update.cox, type = "dfbeta",linear.predictions = FALSE,ggtheme = the
me_minimal())
```

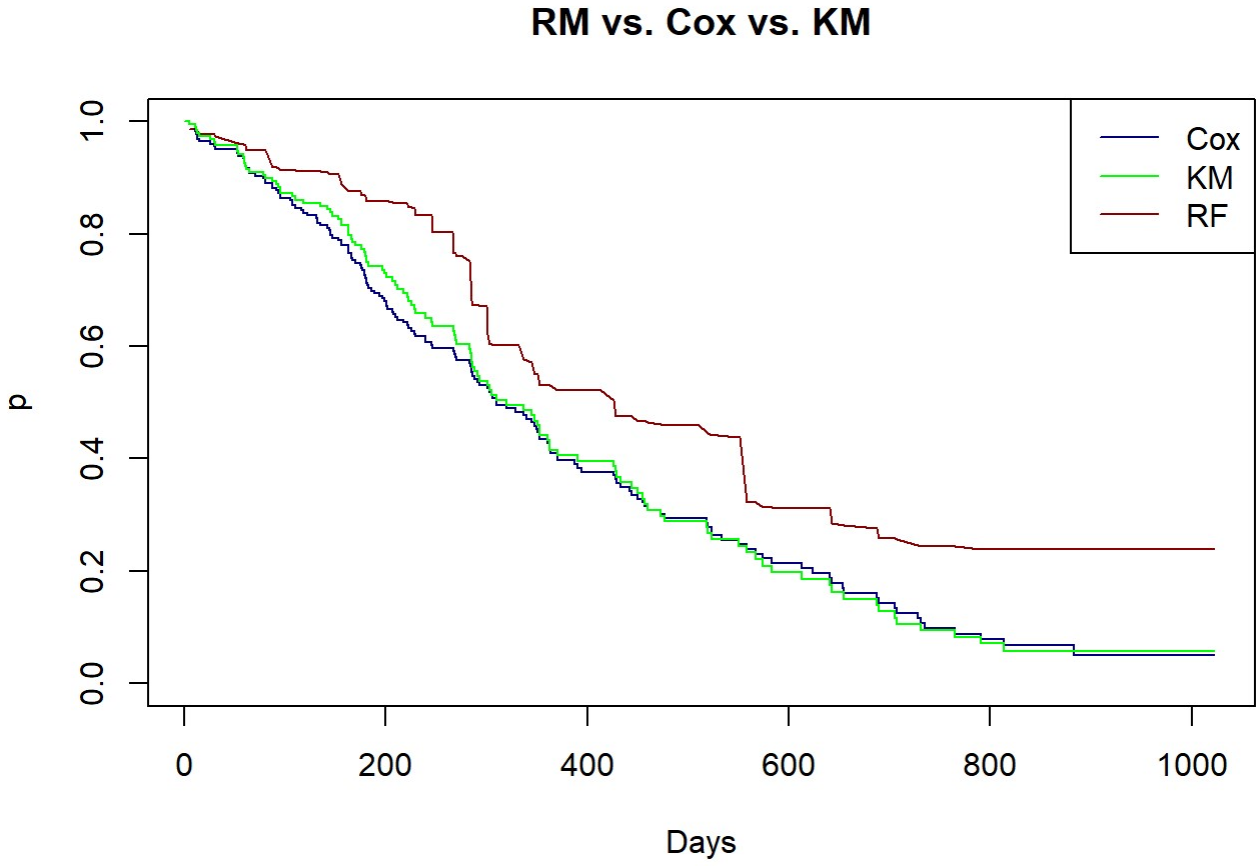



```
#testing for outliers in each of the covarients
```

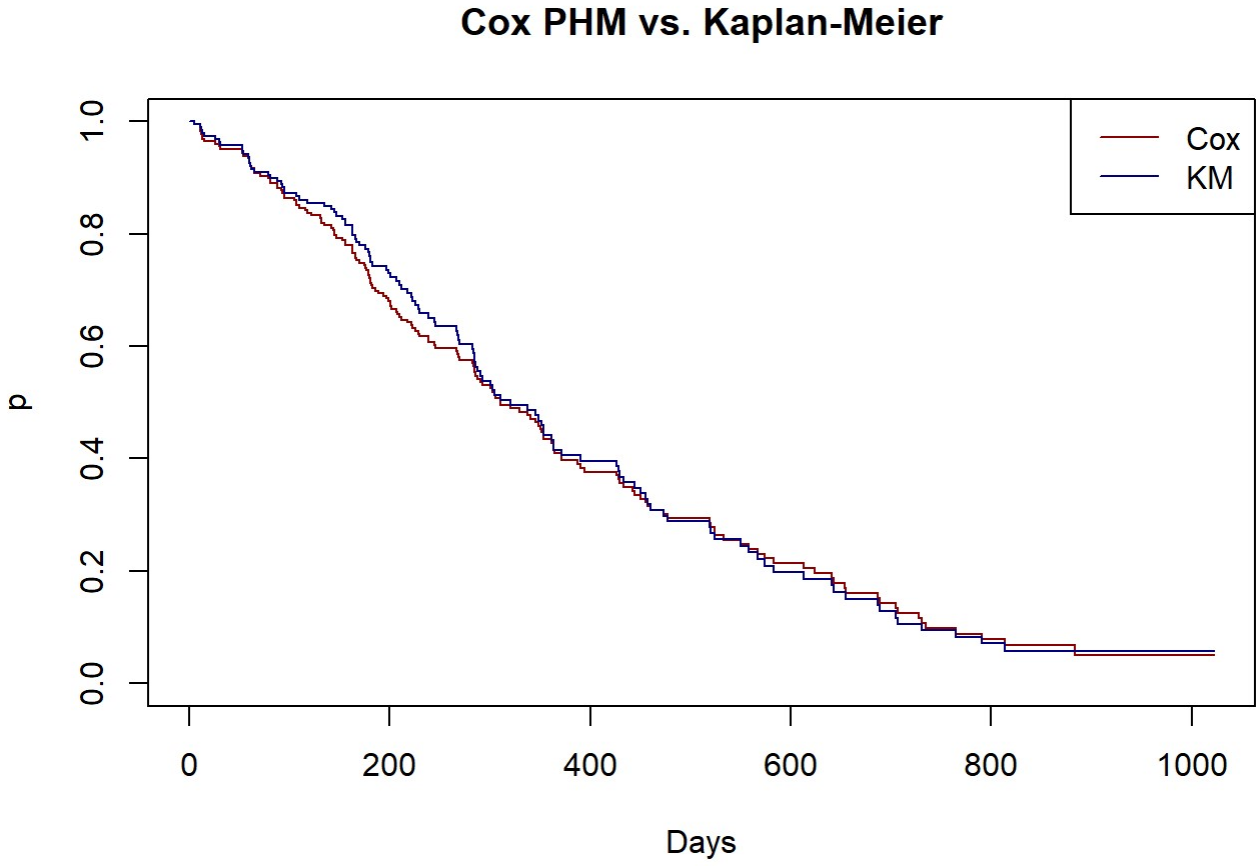
While age was significant when considered on its own after the stepwise process it is exclused from the covarients

```
#install.packages("ranger")
library(ranger)

#Fitting the random forest
ranger <- ranger(Surv(lung$time,lung$status==2) ~.,data=lung,num.trees = 500, importa
nce = "permutation",seed = 1)
```

```
#ploting all models
coxfit <- survfit(update.cox)
rangePlot <- plot(ranger$unique.death.times,ranger$survival[1,], type = "l", ylim = c
(0,1),col = "darkred", main = "RM vs. Cox vs. KM", xlab = "Days", ylab = "p")
lines(all,conf.int = F, col = "navy")
lines(coxfit, col = "green", conf.int = F)
legend("topright",legend = c("Cox", "KM", "RF"), col = c("navy", "green","darkred"),l
ty = 1)
```




```
plot(all, main = "Cox PHM vs. Kaplan-Meier", xlab = "Days", ylab = "p", col = "darkre
d", conf.int = F)
lines(survfit(update.cox),conf.int = F, col = "navy")
legend("topright",legend = c("Cox", "KM"), col = c("darkred", "navy"),lty = 1)
```




```
data.frame(sort(ranger$variable.importance,decreasing = TRUE))
```

```
##           sort.ranger.variable.importance..decreasing...TRUE.
## ph.ecog                                            0.0213567476
## pat.karno                                          0.0184760718
## sex                                                0.0057961043
## wt.loss                                            0.0026184009
## ph.karno                                           0.0022799733
## age                                                0.0001386971
## inst                                              -0.0006621027
## meal.cal                                          -0.0045072053
```

```
#shows the importance of each variable as determined by the random forest model
```