Survival Analysis

#Importing Libraries  
library(readtext)

## Warning: package 'readtext' was built under R version 4.0.4

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.3

## -- Attaching packages ------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.1 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.4

## Warning: package 'stringr' was built under R version 4.0.3

## -- Conflicts ---------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(survival)  
library(survminer)

## Warning: package 'survminer' was built under R version 4.0.4

## Loading required package: ggpubr

## Warning: package 'ggpubr' was built under R version 4.0.3

library(stargazer)

## Warning: package 'stargazer' was built under R version 4.0.3

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

#library(car)  
#library(lmtest)  
#library("ggpubr")  
#library(data.table)  
#library(MASS)  
#library(AER)  
#library(pscl)  
  
#Setting the Working Directory and Importing the Dataset  
setwd("C:/Users/surya/Downloads")  
  
lc <- read.table("LungCancer.txt", skip=15)  
colnames(lc) <- c("treatment", "cell\_type", "survival\_days", "status", "karno\_score",  
 "months\_diag", "age\_years", "pri\_chemo")  
attach(lc)  
  
#NA Values Column-Wise & Setting Factors  
sapply(lc, function(x) sum(is.na(x)))

## treatment cell\_type survival\_days status karno\_score   
## 0 0 0 0 0   
## months\_diag age\_years pri\_chemo   
## 0 0 0

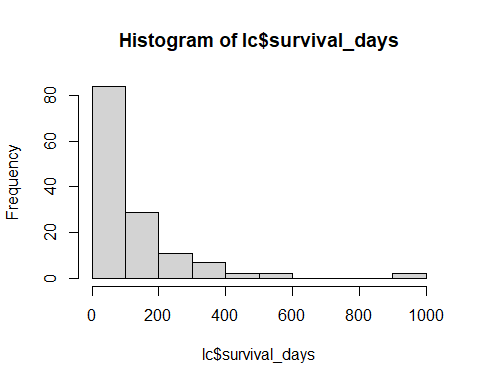
str(lc)

## 'data.frame': 137 obs. of 8 variables:  
## $ treatment : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ cell\_type : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ survival\_days: int 72 411 228 126 118 10 82 110 314 100 ...  
## $ status : int 1 1 1 1 1 1 1 1 1 0 ...  
## $ karno\_score : int 60 70 60 60 70 20 40 80 50 70 ...  
## $ months\_diag : int 7 5 3 9 11 5 10 29 18 6 ...  
## $ age\_years : int 69 64 38 63 65 49 69 68 43 70 ...  
## $ pri\_chemo : int 0 10 0 10 10 0 10 0 0 0 ...

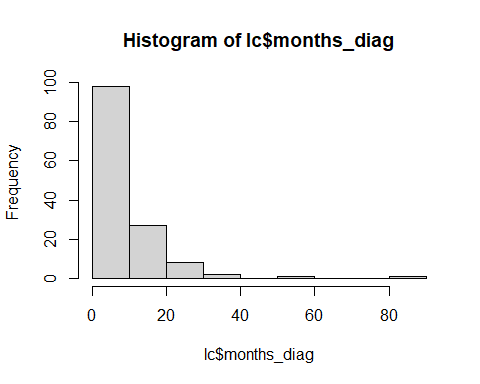
cols <- c(1, 2, 4, 8)  
lc[cols] <- lapply(lc[cols], factor)  
str(lc)

## 'data.frame': 137 obs. of 8 variables:  
## $ treatment : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...  
## $ cell\_type : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...  
## $ survival\_days: int 72 411 228 126 118 10 82 110 314 100 ...  
## $ status : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 1 ...  
## $ karno\_score : int 60 70 60 60 70 20 40 80 50 70 ...  
## $ months\_diag : int 7 5 3 9 11 5 10 29 18 6 ...  
## $ age\_years : int 69 64 38 63 65 49 69 68 43 70 ...  
## $ pri\_chemo : Factor w/ 2 levels "0","10": 1 2 1 2 2 1 2 1 1 1 ...

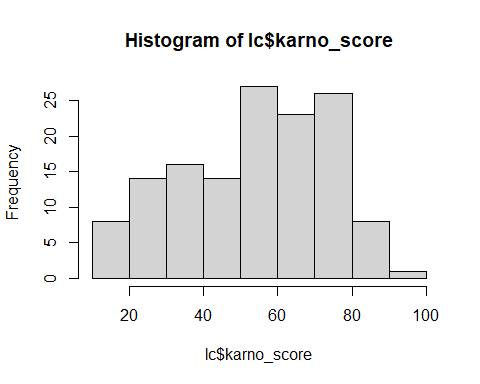
#Data Visualizations and Feature Engineering  
hist(lc$survival\_days)



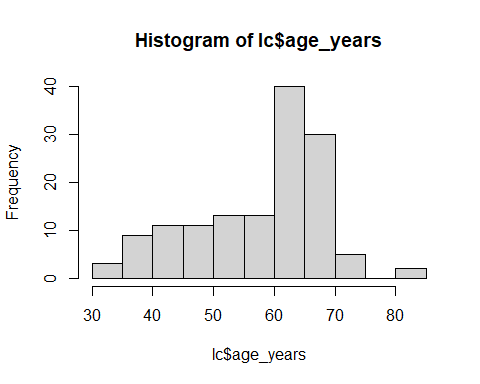
hist(lc$months\_diag)



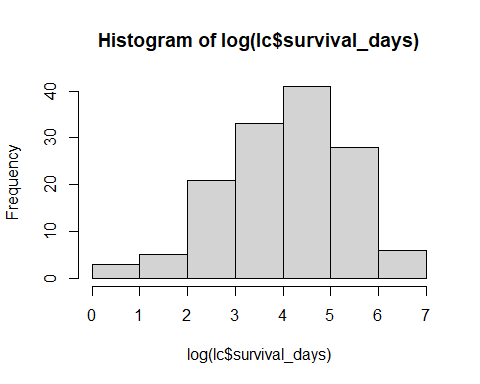
hist(lc$karno\_score)



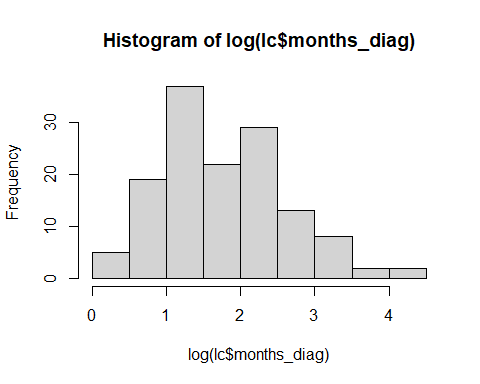
hist(lc$age\_years)



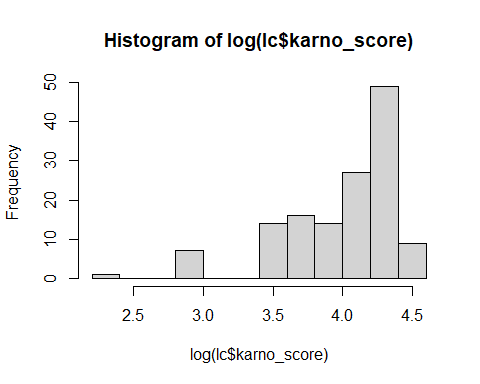
hist(log(lc$survival\_days))



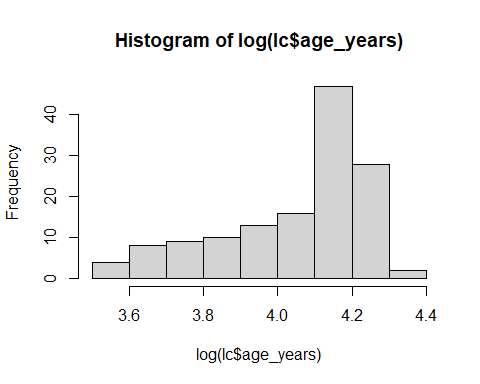
hist(log(lc$months\_diag))



hist(log(lc$karno\_score))



hist(log(lc$age\_years))



table(lc$status)

##   
## 0 1   
## 9 128

table(lc$treatment, lc$cell\_type)

##   
## 1 2 3 4  
## 1 15 30 9 15  
## 2 20 18 18 12

table(lc$cell\_type, lc$pri\_chemo)

##   
## 0 10  
## 1 21 14  
## 2 37 11  
## 3 22 5  
## 4 17 10

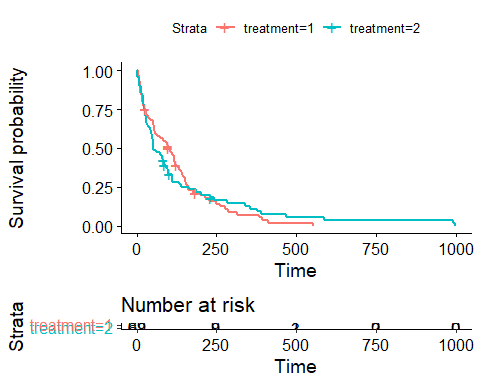
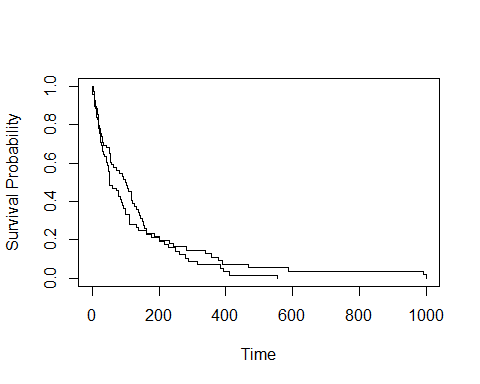
lc$cell\_type = ifelse(lc$cell\_type == 2, "small cell", "non-small cell")  
  
table(cell\_type, pri\_chemo)

## pri\_chemo  
## cell\_type 0 10  
## 1 21 14  
## 2 37 11  
## 3 22 5  
## 4 17 10

#Non-Parametric Models  
#Kaplan-Meier Model  
y <- Surv(survival\_days, status)  
km1 <- survfit(y ~ treatment, data=lc)  
summary(km1)

## Call: survfit(formula = y ~ treatment, data = lc)  
##   
## treatment=1   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 3 69 1 0.9855 0.0144 0.95771 1.000  
## 4 68 1 0.9710 0.0202 0.93223 1.000  
## 7 67 1 0.9565 0.0246 0.90959 1.000  
## 8 66 2 0.9275 0.0312 0.86834 0.991  
## 10 64 2 0.8986 0.0363 0.83006 0.973  
## 11 62 1 0.8841 0.0385 0.81165 0.963  
## 12 61 2 0.8551 0.0424 0.77592 0.942  
## 13 59 1 0.8406 0.0441 0.75849 0.932  
## 16 58 1 0.8261 0.0456 0.74132 0.921  
## 18 57 2 0.7971 0.0484 0.70764 0.898  
## 20 55 1 0.7826 0.0497 0.69109 0.886  
## 21 54 1 0.7681 0.0508 0.67472 0.874  
## 22 53 1 0.7536 0.0519 0.65851 0.862  
## 27 51 1 0.7388 0.0529 0.64208 0.850  
## 30 50 1 0.7241 0.0539 0.62580 0.838  
## 31 49 1 0.7093 0.0548 0.60967 0.825  
## 35 48 1 0.6945 0.0556 0.59368 0.812  
## 42 47 1 0.6797 0.0563 0.57782 0.800  
## 51 46 1 0.6650 0.0570 0.56209 0.787  
## 52 45 1 0.6502 0.0576 0.54649 0.774  
## 54 44 2 0.6206 0.0587 0.51565 0.747  
## 56 42 1 0.6059 0.0591 0.50040 0.734  
## 59 41 1 0.5911 0.0595 0.48526 0.720  
## 63 40 1 0.5763 0.0598 0.47023 0.706  
## 72 39 1 0.5615 0.0601 0.45530 0.693  
## 82 38 1 0.5467 0.0603 0.44049 0.679  
## 92 37 1 0.5320 0.0604 0.42577 0.665  
## 95 36 1 0.5172 0.0605 0.41116 0.651  
## 100 34 1 0.5020 0.0606 0.39615 0.636  
## 103 32 1 0.4863 0.0607 0.38070 0.621  
## 105 31 1 0.4706 0.0608 0.36537 0.606  
## 110 30 1 0.4549 0.0607 0.35018 0.591  
## 117 29 2 0.4235 0.0605 0.32017 0.560  
## 118 27 1 0.4079 0.0602 0.30537 0.545  
## 122 26 1 0.3922 0.0599 0.29069 0.529  
## 126 24 1 0.3758 0.0596 0.27542 0.513  
## 132 23 1 0.3595 0.0592 0.26031 0.496  
## 139 22 1 0.3432 0.0587 0.24535 0.480  
## 143 21 1 0.3268 0.0582 0.23057 0.463  
## 144 20 1 0.3105 0.0575 0.21595 0.446  
## 151 19 1 0.2941 0.0568 0.20151 0.429  
## 153 18 1 0.2778 0.0559 0.18725 0.412  
## 156 17 1 0.2614 0.0550 0.17317 0.395  
## 162 16 2 0.2288 0.0527 0.14563 0.359  
## 177 14 1 0.2124 0.0514 0.13218 0.341  
## 200 12 1 0.1947 0.0501 0.11761 0.322  
## 216 11 1 0.1770 0.0486 0.10340 0.303  
## 228 10 1 0.1593 0.0468 0.08956 0.283  
## 250 9 1 0.1416 0.0448 0.07614 0.263  
## 260 8 1 0.1239 0.0426 0.06318 0.243  
## 278 7 1 0.1062 0.0400 0.05076 0.222  
## 287 6 1 0.0885 0.0371 0.03896 0.201  
## 314 5 1 0.0708 0.0336 0.02793 0.180  
## 384 4 1 0.0531 0.0295 0.01788 0.158  
## 392 3 1 0.0354 0.0244 0.00917 0.137  
## 411 2 1 0.0177 0.0175 0.00256 0.123  
## 553 1 1 0.0000 NaN NA NA  
##   
## treatment=2   
## time n.risk n.event survival std.err lower 95% CI upper 95% CI  
## 1 68 2 0.9706 0.0205 0.93125 1.000  
## 2 66 1 0.9559 0.0249 0.90830 1.000  
## 7 65 2 0.9265 0.0317 0.86647 0.991  
## 8 63 2 0.8971 0.0369 0.82766 0.972  
## 13 61 1 0.8824 0.0391 0.80900 0.962  
## 15 60 2 0.8529 0.0429 0.77278 0.941  
## 18 58 1 0.8382 0.0447 0.75513 0.930  
## 19 57 2 0.8088 0.0477 0.72056 0.908  
## 20 55 1 0.7941 0.0490 0.70360 0.896  
## 21 54 1 0.7794 0.0503 0.68684 0.884  
## 24 53 2 0.7500 0.0525 0.65383 0.860  
## 25 51 3 0.7059 0.0553 0.60548 0.823  
## 29 48 1 0.6912 0.0560 0.58964 0.810  
## 30 47 1 0.6765 0.0567 0.57394 0.797  
## 31 46 1 0.6618 0.0574 0.55835 0.784  
## 33 45 1 0.6471 0.0580 0.54289 0.771  
## 36 44 1 0.6324 0.0585 0.52754 0.758  
## 43 43 1 0.6176 0.0589 0.51230 0.745  
## 44 42 1 0.6029 0.0593 0.49717 0.731  
## 45 41 1 0.5882 0.0597 0.48216 0.718  
## 48 40 1 0.5735 0.0600 0.46724 0.704  
## 49 39 1 0.5588 0.0602 0.45244 0.690  
## 51 38 2 0.5294 0.0605 0.42313 0.662  
## 52 36 2 0.5000 0.0606 0.39423 0.634  
## 53 34 1 0.4853 0.0606 0.37993 0.620  
## 61 33 1 0.4706 0.0605 0.36573 0.606  
## 73 32 1 0.4559 0.0604 0.35163 0.591  
## 80 31 2 0.4265 0.0600 0.32373 0.562  
## 84 28 1 0.4112 0.0597 0.30935 0.547  
## 87 27 1 0.3960 0.0594 0.29509 0.531  
## 90 25 1 0.3802 0.0591 0.28028 0.516  
## 95 24 1 0.3643 0.0587 0.26560 0.500  
## 99 23 2 0.3326 0.0578 0.23670 0.467  
## 111 20 2 0.2994 0.0566 0.20673 0.434  
## 112 18 1 0.2827 0.0558 0.19203 0.416  
## 133 17 1 0.2661 0.0550 0.17754 0.399  
## 140 16 1 0.2495 0.0540 0.16326 0.381  
## 164 15 1 0.2329 0.0529 0.14920 0.363  
## 186 14 1 0.2162 0.0517 0.13538 0.345  
## 201 13 1 0.1996 0.0503 0.12181 0.327  
## 231 12 1 0.1830 0.0488 0.10851 0.308  
## 242 10 1 0.1647 0.0472 0.09389 0.289  
## 283 9 1 0.1464 0.0454 0.07973 0.269  
## 340 8 1 0.1281 0.0432 0.06609 0.248  
## 357 7 1 0.1098 0.0407 0.05304 0.227  
## 378 6 1 0.0915 0.0378 0.04067 0.206  
## 389 5 1 0.0732 0.0344 0.02912 0.184  
## 467 4 1 0.0549 0.0303 0.01861 0.162  
## 587 3 1 0.0366 0.0251 0.00953 0.140  
## 991 2 1 0.0183 0.0180 0.00265 0.126  
## 999 1 1 0.0000 NaN NA NA

plot(km1, xlab="Time", ylab="Survival Probability")  
ggsurvplot(km1, data = lc, risk.table = TRUE)



#For 365 Days  
summary(km1, times=365)

## Call: survfit(formula = y ~ treatment, data = lc)  
##   
## treatment=1   
## time n.risk n.event survival std.err lower 95% CI   
## 365.0000 4.0000 60.0000 0.0708 0.0336 0.0279   
## upper 95% CI   
## 0.1795   
##   
## treatment=2   
## time n.risk n.event survival std.err lower 95% CI   
## 365.0000 6.0000 58.0000 0.1098 0.0407 0.0530   
## upper 95% CI   
## 0.2272

#For 183 Days  
summary(km1, times=183)

## Call: survfit(formula = y ~ treatment, data = lc)  
##   
## treatment=1   
## time n.risk n.event survival std.err lower 95% CI   
## 183.0000 12.0000 52.0000 0.2124 0.0514 0.1322   
## upper 95% CI   
## 0.3414   
##   
## treatment=2   
## time n.risk n.event survival std.err lower 95% CI   
## 183.0000 14.0000 51.0000 0.2329 0.0529 0.1492   
## upper 95% CI   
## 0.3634

#Mean Survival Days  
print(km1, print.mean=TRUE)

## Call: survfit(formula = y ~ treatment, data = lc)  
##   
## n events median 0.95LCL 0.95UCL  
## treatment=1 69 64 103.0 59 132  
## treatment=2 68 64 52.5 44 95

#Cox Proportional Hazard Model  
cox <- coxph(y ~ treatment + cell\_type + months\_diag + age\_years +  
 pri\_chemo + treatment\*cell\_type + treatment\*pri\_chemo + treatment\*age\_years, data = lc, method="breslow")  
  
#Parametric Models  
#Exponential, Weibull, and log-logistic Models  
exp <- survreg(y ~ treatment + cell\_type + months\_diag + age\_years +  
 pri\_chemo + treatment\*cell\_type + treatment\*pri\_chemo, data = lc, dist="exponential")  
  
wbl <- survreg(y ~ treatment + cell\_type + months\_diag + age\_years +   
 pri\_chemo + treatment\*cell\_type + treatment\*pri\_chemo, data = lc, dist="weibull")  
  
llg <- survreg(y ~ treatment + cell\_type + months\_diag + age\_years +   
 pri\_chemo + treatment\*cell\_type + treatment\*pri\_chemo, data = lc, dist="loglogistic")  
  
#Stargazer  
stargazer(cox, exp, wbl, llg, type="text", single.row=TRUE)

##   
## =======================================================================================================  
## Dependent variable:   
## ------------------------------------------------------------------------  
## y   
## Cox exponential Weibull survreg: loglogistic  
## prop. hazards   
## (1) (2) (3) (4)   
## -------------------------------------------------------------------------------------------------------  
## treatment2 0.595 (1.258) -0.199 (0.269) -0.222 (0.288) -0.568\* (0.316)   
## cell\_typesmall cell 0.415 (0.260) -0.456\* (0.257) -0.474\* (0.275) -0.835\*\*\* (0.315)   
## months\_diag 0.011 (0.010) -0.012 (0.010) -0.012 (0.010) -0.010 (0.012)   
## age\_years 0.002 (0.012) -0.002 (0.010) -0.001 (0.011) 0.005 (0.011)   
## pri\_chemo10 0.151 (0.280) -0.168 (0.279) -0.183 (0.298) -0.418 (0.354)   
## treatment2:cell\_typesmall cell 0.502 (0.401) -0.556 (0.386) -0.522 (0.414) 0.055 (0.480)   
## treatment2:pri\_chemo10 -0.771\* (0.438) 0.862\*\* (0.421) 0.875\* (0.449) 0.799 (0.533)   
## treatment2:age\_years -0.006 (0.021)   
## Constant 5.246\*\*\* (0.578) 5.214\*\*\* (0.615) 4.580\*\*\* (0.666)   
## -------------------------------------------------------------------------------------------------------  
## Observations 137 137 137 137   
## R2 0.113   
## Max. Possible R2 0.999   
## Log Likelihood -497.694 -738.964 -738.475 -742.785   
## chi2 (df = 7) 24.514\*\*\* 19.233\*\*\* 14.961\*\*   
## Wald Test 17.620\*\* (df = 8)   
## LR Test 16.381\*\* (df = 8)   
## Score (Logrank) Test 18.635\*\* (df = 8)   
## =======================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

#Cox Proportional Hazard Model  
cox1 <- coxph(y ~ treatment + cell\_type + months\_diag + age\_years +  
 pri\_chemo + treatment\*age\_years + treatment\*months\_diag, data = lc, method="breslow")  
  
#Parametric Models  
#Exponential, Weibull, and log-logistic Models  
exp1 <- survreg(y ~ treatment + cell\_type + months\_diag + age\_years +  
 pri\_chemo + treatment\*age\_years + treatment\*months\_diag, data = lc, dist="exponential")  
  
wbl1 <- survreg(y ~ treatment + cell\_type + months\_diag + age\_years +   
 pri\_chemo + treatment\*age\_years + treatment\*months\_diag, data = lc, dist="weibull")  
  
llg1 <- survreg(y ~ treatment + cell\_type + months\_diag + age\_years +   
 pri\_chemo + treatment\*age\_years + treatment\*months\_diag, data = lc, dist="loglogistic")  
  
#Stargazer  
stargazer(cox1, exp1, wbl1, llg1, type="text", single.row=TRUE)

##   
## ================================================================================================  
## Dependent variable:   
## -------------------------------------------------------------------------  
## y   
## Cox exponential Weibull survreg: loglogistic  
## prop. hazards   
## (1) (2) (3) (4)   
## ------------------------------------------------------------------------------------------------  
## treatment2 -0.057 (1.251) 0.317 (1.287) 0.233 (1.388) -1.678 (1.303)   
## cell\_typesmall cell 0.610\*\*\* (0.207) -0.688\*\*\* (0.205) -0.698\*\*\* (0.223) -0.852\*\*\* (0.240)   
## months\_diag 0.010 (0.017) -0.009 (0.017) -0.010 (0.019) -0.018 (0.021)   
## age\_years 0.002 (0.012) -0.005 (0.012) -0.004 (0.013) -0.004 (0.014)   
## pri\_chemo10 -0.171 (0.232) 0.247 (0.232) 0.229 (0.253) -0.070 (0.278)   
## treatment2:age\_years 0.004 (0.021) -0.007 (0.021) -0.006 (0.023) 0.020 (0.021)   
## treatment2:months\_diag -0.002 (0.019) -0.002 (0.019) 0.0001 (0.021) 0.017 (0.024)   
## Constant 5.409\*\*\* (0.703) 5.357\*\*\* (0.775) 5.113\*\*\* (0.847)   
## ------------------------------------------------------------------------------------------------  
## Observations 137 137 137 137   
## R2 0.079   
## Max. Possible R2 0.999   
## Log Likelihood -500.222 -742.473 -741.405 -743.262   
## chi2 (df = 7) 17.496\*\* 13.373\* 14.007\*   
## Wald Test 12.000 (df = 7)   
## LR Test 11.324 (df = 7)   
## Score (Logrank) Test 12.329\* (df = 7)   
## ================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01