Coursework - Survival models

CID: 01708919

1.

$$\begin{aligned} P(T > 15) = S(15) &= \prod_{j:a_j \leq 15} (1 - h_j) \\ &= \prod_{j=1}^{3} (1 - h_j) \\ &= 0.75^2 * 0.5 \\ &= \frac{9}{32} \end{aligned}$$

2.

Firstly load the dataset.

```
data <- read.csv("C:\\Users\\pc\\Downloads\\cwdat.csv")
```

(a)

The parameter θ is given by the implementation below.

```
model_exp <- list(
  validtheta= function(theta) theta>0,
  h=function(x,theta) rep(theta,length(x)),
  H=function(x,theta) theta*x
)

l_exp <- function(theta,data) {
  if (!model_exp$validtheta(theta)) return(-Inf)
  sum(log(model_exp$h(data$T[data$Delta==1],theta)))-sum(model_exp$H(data$T,theta))
}

o_exp <- optim(c(1),fn=function(theta) -l_exp(theta,data),
      method="Brent",lower=1e-4,upper=1e6,
      hessian=TRUE)

print(paste('theta:',o_exp$par))
```

```
## [1] "theta: 0.369333551574242"
```

(b)

The parameters α, η are given by the implementation below.

```
model_wei <- list(
  validalpha= function(alpha) alpha>0,
  valideta= function(eta) eta>0,
  h=function(x,alpha,eta) eta*alpha**(-eta)*x**(eta-1),
  H=function(x,alpha,eta) (x/alpha)**eta
)

l_wei <- function(theta,data) {
  alpha <- theta[1]
  eta <- theta[2]
  if (!model_wei$validalpha(alpha)) return(-Inf)
  if (!model_wei$valideta(eta)) return(-Inf)
  sum(log(model_wei$h(data$T[data$Delta==1],alpha,eta)))-sum(model_wei$H(data$T,alpha,eta))
}

o_wei <- optim(c(1,1),fn=function(theta) -l_wei(theta,data),
      method="BFGS",hessian=TRUE)

print(paste('alpha:',o_wei$par[1]))
```

```
## [1] "alpha: 2.29082531436014"
```

```
print(paste('eta:',o_wei$par[2]))
```

```
## [1] "eta: 0.624193010356495"
```

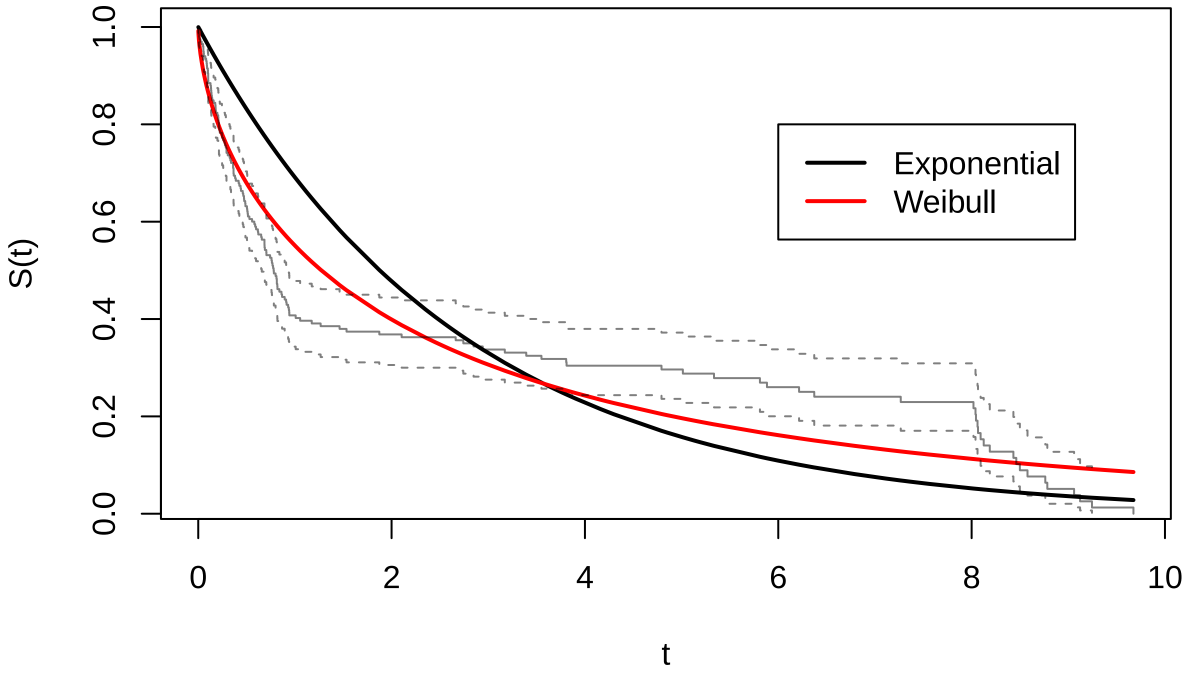
(c)

The plot of both distributions and the Kaplan-Meier estimate is given by the implentation below.

```
library(survival)
S_exp <- function(x,theta) exp(-theta*x)
S_wei <- function(x,theta){
  alpha <- theta[1]
  eta <- theta[2]
  exp(-(x/alpha)**eta)
}
T_sorted <- sort(data$T)
plot(T_sorted,sapply(T_sorted,function(y) S_exp(y,o_exp$par)),
     type="l",
     lwd=2,
     ylab=expression("S(t)"),
     xlab=expression("t"))

lines(T_sorted,sapply(T_sorted,function(y) S_wei(y,o_wei$par)),
      lwd=2,col="red")
legend(6,0.8,c("Exponential","Weibull"),lwd=c(2,2),col=c("black","red"))
```

```
fit <- survfit(Surv(data$T,data$Delta,type='right')~1)
lines(fit,col=rgb(0,0,0,0.5))
```

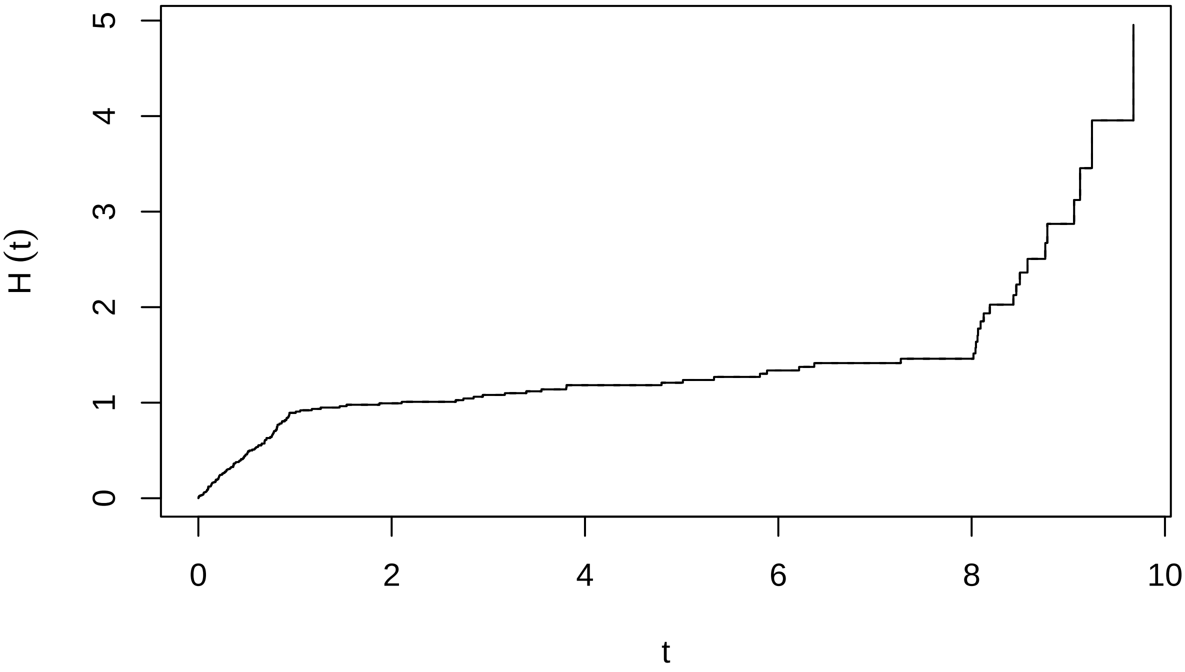


It is clear from the figure that the model of Weibull distribution is closer to the Kaplan-Meier estimate of the data set. Hence Weibull distribution models the data more appropriately than exponential distribution. However, Weibull distribution may not be considered as an adequate choice of parametric distribution as well since its values of survival functions for $t \in [4, 8]$ also lie outside the 95 confidence interval (although much closer than the exponential model.)

(d)

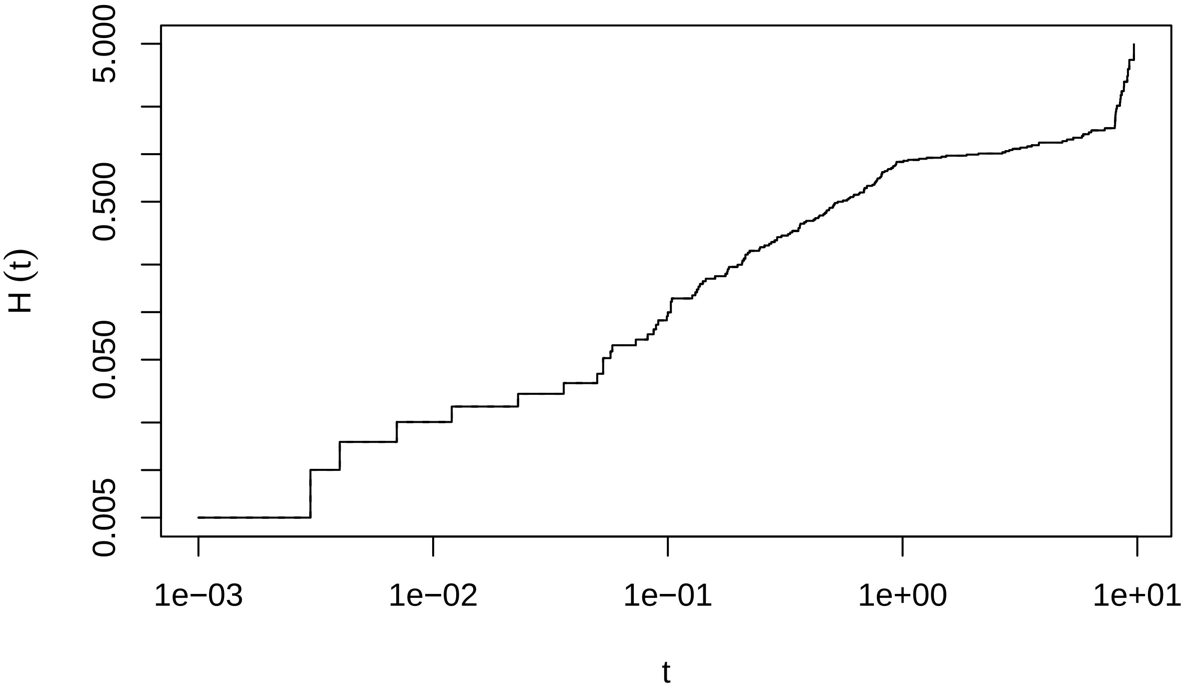
Here the estimates of cumulative hazard function $\hat{H}(t)$ are plotted using the $plot.survfit$ function similarly in *(d)* but setting the argument $fun =' cumhaz'$ to plot the cumulative hazard curve of the Kaplan-Meier estimate (essentially equivalent to the Nelson-Aalen estimate).

```
plot(fit,fun="cumhaz",xlab="t",ylab=expression(H~(t)),conf.int=0)
```



We can observe that the cumulative hazard is clearly not linear with time, hence exponential distribution is inappropriate. Then we produce the log-log plot. Note that changing to logarithmic scale is equivalent to taking log of plotted values.

```
plot(fit,fun="cumhaz",xlab="t",ylab=expression(H~(t)),log="xy",conf.int=0)
```



We can see that a linear correlation of $\log \hat{H}(t)$ vs. $\log t$ is more adequate. This supports the conclusion above that Weibull distribution is a more appropriate model.

3.

(a)

The data set we will use for this part would be the *gbsg* data set from the *survival* package. The data set contains observations of 686 patients from a 1984-1989 trial conducted by the German Breast Cancer Study Group (GBSG) of 720 patients with node positive breast cancer. The 5 main covariates that will be of interest are *age*: of patients in years; *meno*: menopausal status (0= premenopausal, 1= postmenopausal); *grade*: tumor grade; *nodes*:number of positive lymph nodes; *hormon*: hormonal therapy (0= no, 1= yes).

(b)

We will fit a Cox proportional hazards regression model as the semi-parametric model.

```
library(survival)

fitcox1 <- coxph(Surv(rfstime,status)~age,data=gbsg)
fitcox1
```

```
## Call:
## coxph(formula = Surv(rfstime, status) ~ age, data = gbsg)
```

```
##
##          coef exp(coef)  se(coef)      z     p
## age -0.004485  0.995525  0.005887 -0.762 0.446
##
## Likelihood ratio test=0.58  on 1 df, p=0.4462
## n= 686, number of events= 299
```

```
fitcox2 <- coxph(Surv(rfstime,status)~factor(meno),data=gbsg)
fitcox2
```

```
## Call:
## coxph(formula = Surv(rfstime, status) ~ factor(meno), data = gbsg)
##
##                  coef exp(coef) se(coef)    z     p
## factor(meno)1 0.06265   1.06466  0.11824 0.53 0.596
##
## Likelihood ratio test=0.28  on 1 df, p=0.5955
## n= 686, number of events= 299
```

```
fitcox3 <- coxph(Surv(rfstime,status)~grade,data=gbsg)
fitcox3
```

```
## Call:
## coxph(formula = Surv(rfstime, status) ~ grade, data = gbsg)
##
##         coef exp(coef) se(coef)     z        p
## grade 0.4490    1.5667   0.1006 4.464 8.03e-06
##
## Likelihood ratio test=19.99  on 1 df, p=7.803e-06
## n= 686, number of events= 299
```

```
fitcox4 <- coxph(Surv(rfstime,status)~nodes,data=gbsg)
fitcox4
```

```
## Call:
## coxph(formula = Surv(rfstime, status) ~ nodes, data = gbsg)
##
##          coef exp(coef) se(coef)     z      p
## nodes 0.05860   1.06035  0.00674 8.694 <2e-16
##
## Likelihood ratio test=50.04  on 1 df, p=1.51e-12
## n= 686, number of events= 299
```

```
fitcox5 <- coxph(Surv(rfstime,status)~factor(hormon),data=gbsg)
fitcox5
```

```
## Call:
## coxph(formula = Surv(rfstime, status) ~ factor(hormon), data = gbsg)
##
##                    coef exp(coef) se(coef)      z      p
## factor(hormon)1 -0.3640    0.6949   0.1250 -2.911 0.0036
##
## Likelihood ratio test=8.82  on 1 df, p=0.002977
## n= 686, number of events= 299
```

By analyzing the covariates separately, we can observe that tumor grade and whether having taken hormonal therapy have a significant effect on hazards of survival models. the positive sign of coefficient of tumor grade indicates that higher tumor grade will result in higher risk of death, whereas the coefficient of *hormon* implies that patients who have taken the hormonal therapy would have lower risk of death.

(c)

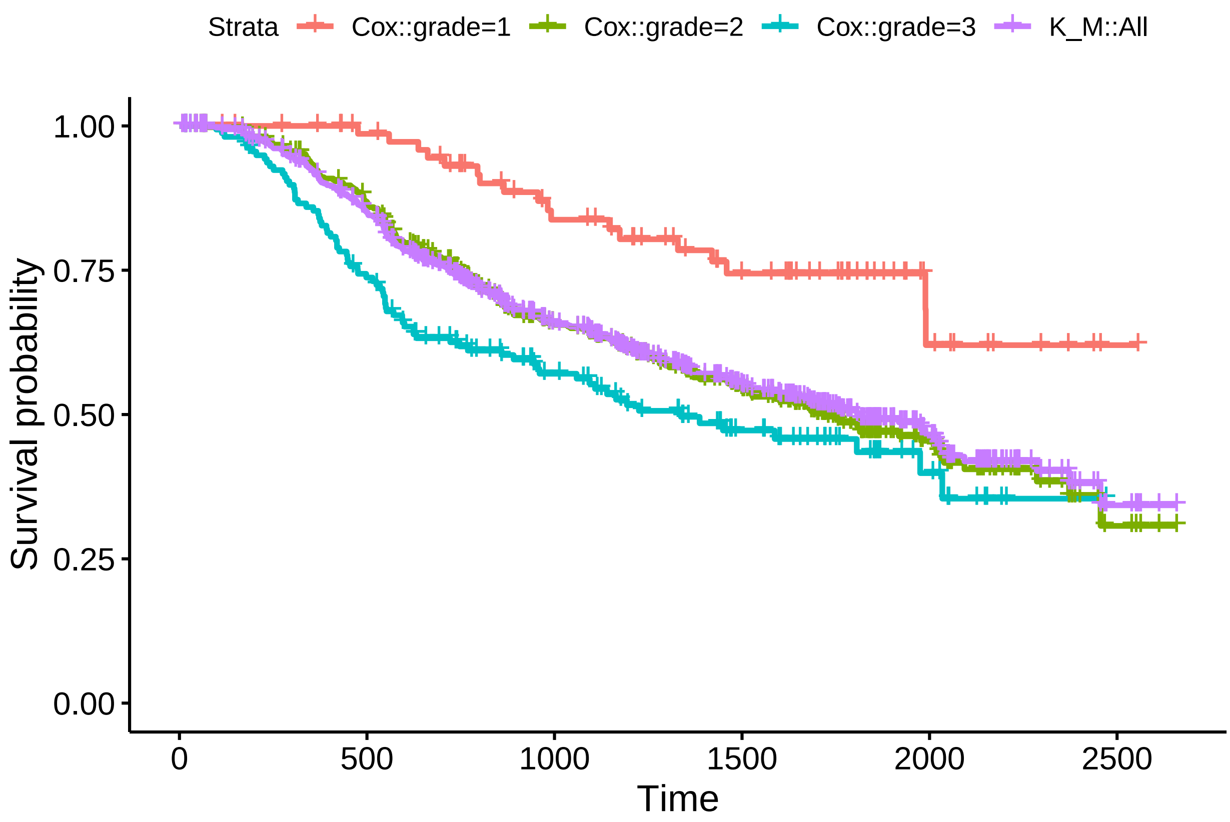
Effect of tumor grade:

```
library(survival)

cox1 <- survfit(Surv(rfstime,status)~grade,data=gbsg)
kaplan <- survfit(Surv(rfstime,status)~1,data=gbsg)

library(survminer)

fit_both <- list(Cox=cox1, K_M=kaplan)
ggsurvplot(fit_both,data=gbsg,combine=TRUE)
```



It is clear that higher tumor grade reduces the chance of survival. Note that the survival probability curve at tumor grade=2 is very close to the K-M estimate (close to baseline).

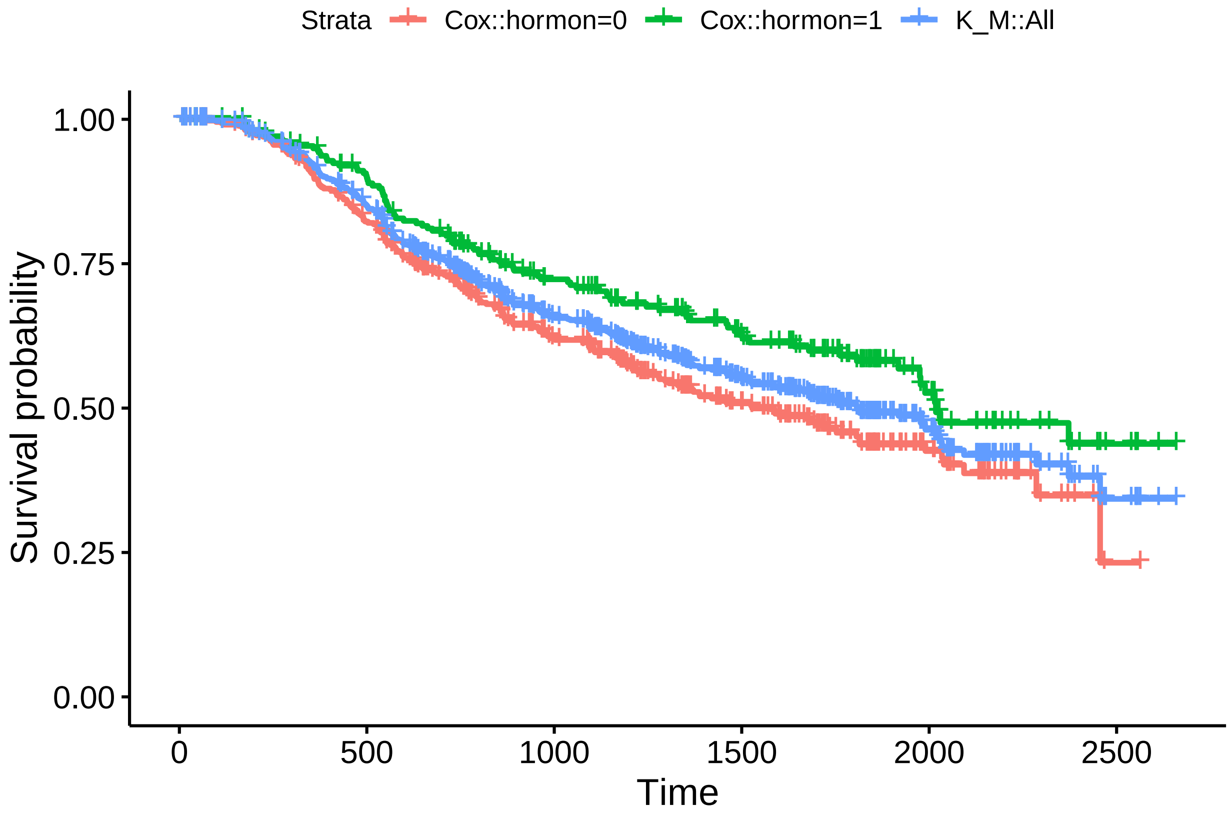
Effect of hormonal therapy:

```
library(survival)

cox2 <- survfit(Surv(rfstime,status)~hormon,data=gbsg)

library(survminer)

fit_both <- list(Cox=cox2, K_M=kaplan)
ggsurvplot(fit_both,data=gbsg,combine=TRUE)
```



We can again observe that patients having taken the hormonal therapy will have higher chance of survival.

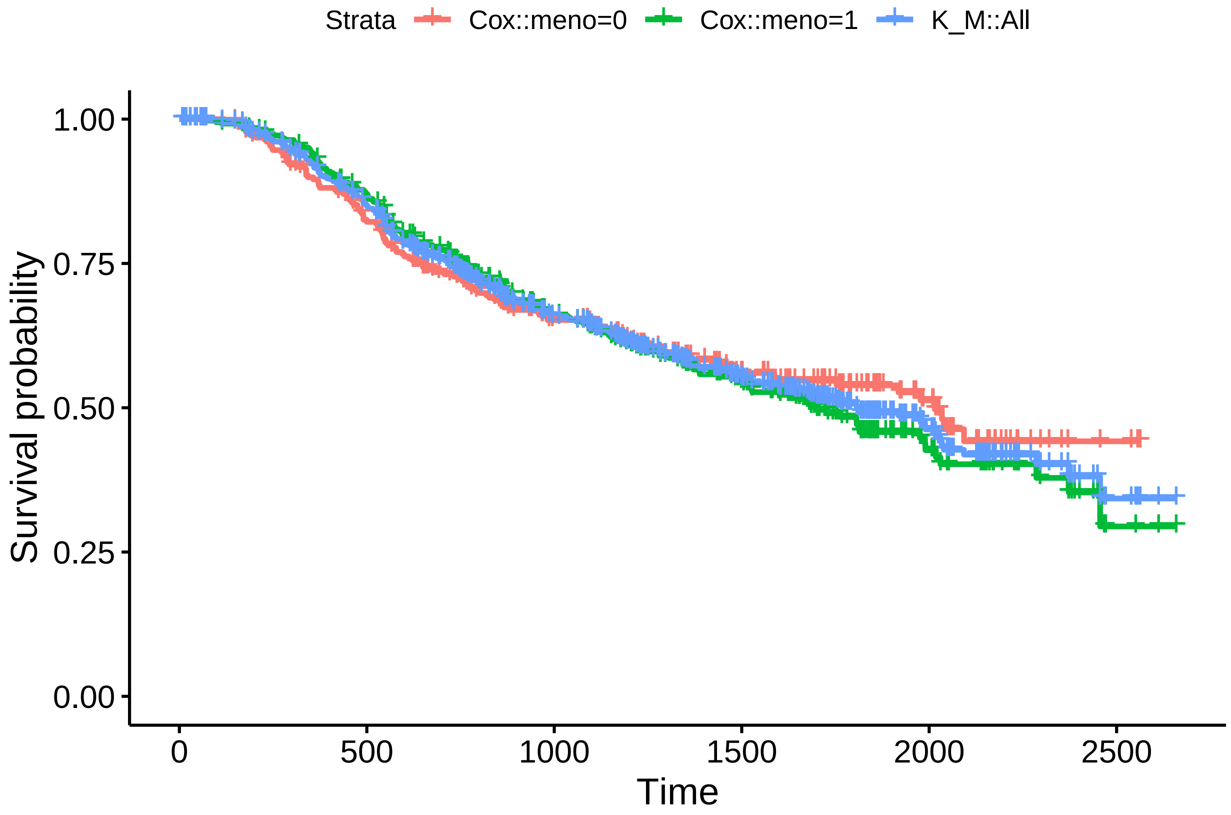
Effect of menopausal status:

```
library(survival)

cox3 <- survfit(Surv(rfstime,status)~meno,data=gbsg)

library(survminer)

fit_both <- list(Cox=cox3, K_M=kaplan)
ggsurvplot(fit_both,data=gbsg,combine=TRUE)
```



We can see that the menopausal status has no significant effect on the survival probabilities.