Survival Analysis

Vazgen Tadevosyan

You are required to submit both Markdown and HTML files with your answers and code in it. Do not remove problems from your Markdown file. Please write your name at the beginning of your file into the YAML-header inside of author.

Data set relevant for this assignment can be found on Moodle. The description of the variables is given in a separate file.

Good luck!

```
                                          Task 1   (4 pt.)
a. Clean and describe the final data. Which of variables can be useful for survival analysis? Why? Which of them
can be an event indicator?
b. What is the difference between interval censored and point censored data?
c. What are the mail causes of censored data (right-censored)? Why  cannot we ignore censoring, bring and exampl
e?
```

```
library(ggplot2)
library(survival)
library(survminer)
df<-read.csv("survdata.csv")

str(df)
```

```
## 'data.frame':    3400 obs. of  15 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ chldage  : num  12 12 3 2 4 12 7 3 7 12 ...
##  $ hospital : int  0 0 0 0 0 0 0 0 0 NA ...
##  $ mthage   : int  22 20 24 22 21 20 24 24 26 21 ...
##  $ alcohol  : int  0 1 3 2 1 0 0 3 2 1 ...
##  $ smoke    : int  0 0 0 2 2 0 0 0 2 0 ...
##  $ region   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ poverty  : int  1 1 1 NA 1 1 1 1 1 1 ...
##  $ bweight  : int  1 0 0 0 1 0 0 0 0 0 ...
```

```
##  $ race     : int  1 1 1 1 1 1 1 1 1 ...
##  $ education: int  10 12 12 9 12 12 12 14 12 12 ...
##  $ nsibs    : int  1 1 2 0 0 0 1 0 0 0 ...
##  $ wmonth   : int  1 2 1 0 0 0 0 4 1 3 ...
##  $ sfmonth  : int  1 2 0 0 0 0 0 2 1 2 ...
##  $ agepn    : int  1 12 3 2 4 12 7 3 6 12 ...
```

```
df[duplicated(df$studentid),][1]
```

```
## [1] X
## <0 rows> (or 0-length row.names)
```

```
df$X<-NULL
#sapply(df, function(x) sum(is.na(x)))
df<-na.omit(df)
df$bweight<-ifelse(df$bweight==1,"Yes","No")
factors<-c("alcohol","smoke","region","poverty","race","bweight")
df$smoke<-ifelse(df$smoke==0,0,1)
df$alcohol<-ifelse(df$alcohol==0,0,1)
df[factors] <- lapply(df[factors], factor)
```

a. Indicator hospitalization can be event indicator and with Age Age child had disease can be our survival object. Other variables such as wmonth,sfmonth,agepn, smoke,alcohol can also be useful for our survival analysis as they can have an impact on indicator for hospitalization.

b.

- Interval censored point - we don't know exactly when some babies were dropped out of study(or lost to be followed or…) for various reasons but we know it was within some interval of time.
- point censored data means we know exactly when data point was censored.

c.

- Suppose you're conducting a study on pregnancy duration. You're ready to complete the study and run your analysis, but some women in the study are still pregnant, so you don't know exactly how long their pregnancies will last. These observations would be right-censored. The "failure," or birth in this case, will occur after the recorded time. It is usually the result of limited resources or competing failure. It is important

to include the censored observations in our analysis because the fact that these items have not yet failed has a big impact on your reliability estimates. For example, when we want to estimate probability of a person to die from cancer given years of illness, we should not ignore censored data points at that age meaning the people are lost to be followed or did not die at that time because of cancer.

```
                                                Task 2   (6 pt.)
  Create survival object to solve the task, use the KM method.

  a. What is the probability of surviving to a certain point in time (baseline model without predictors)? Use n.r
isk and n.event of summary function to comment on this task.
  b. Why is the KM model called non-parametric?
  c. Plot the survival probability and cumulative hazard functions. Comment on it.
  d. Add independent variable/s (with more than 2 categories) to your model to create survival curves for differe
nt groups.
  e. Find the variable with significant differences in survival probabilities. Show the survival plot and add a r
isk table to the graph. Comment on it.
```

```
surv_obj<-Surv(time = df$chldage,event = df$hospital)
km<-survfit(surv_obj~1,data=df)
summary(km)
```

```
## Call: survfit(formula = surv_obj ~ 1, data = df)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1   3302      21    0.994 0.00138        0.991        0.996
##     2   3200      14    0.989 0.00180        0.986        0.993
##     3   3105      12    0.985 0.00210        0.981        0.990
##     4   3010       4    0.984 0.00220        0.980        0.988
##     5   2916       5    0.982 0.00232        0.978        0.987
##     6   2805       5    0.981 0.00245        0.976        0.986
##     7   2706       1    0.980 0.00247        0.976        0.985
##     8   2609       2    0.980 0.00253        0.975        0.985
##     9   2513       4    0.978 0.00264        0.973        0.983
```

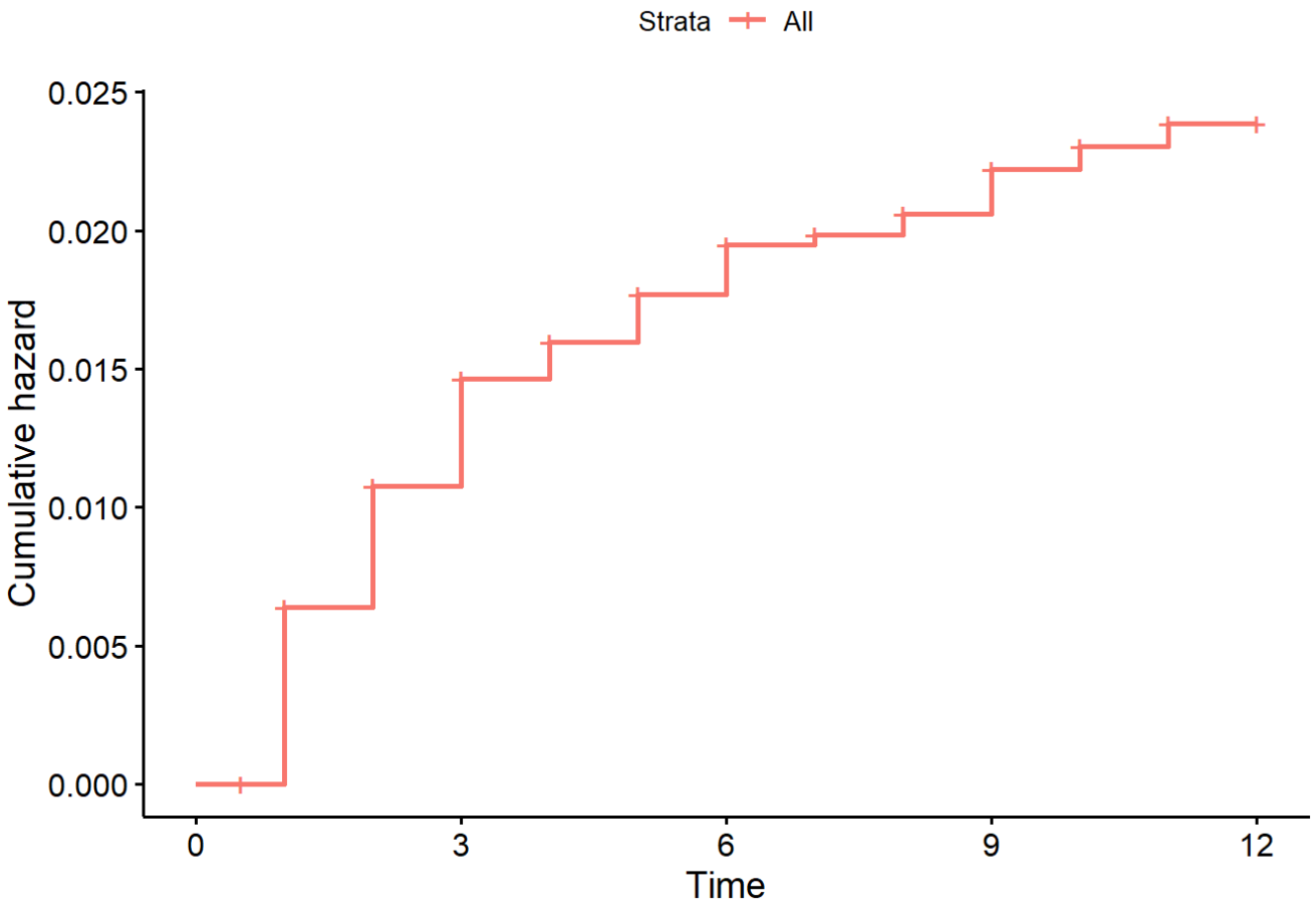
```
##    10   2426       2    0.977 0.00270        0.972        0.983
##    11   2348       2    0.976 0.00276        0.971        0.982
```

So, when a baby has disease for a month the probability to not be hospitalized is or to survive is the following. $p(1) = 1 - \frac{21}{3302} = 0.994$
Probability of a baby to survive 2 months is the following. $p(2) = (1 - \frac{14}{3200}) * p(1) = 0.989$

b. Parametric tests assume underlying statistical distributions in the data. Therefore, several conditions of validity must be met so that the result of a parametric test is reliable. For example assumption about normally distributed data. Non parametric tests do not rely on any distribution. They can thus be applied even if parametric conditions of validity are not met. And When no truncation or censoring occurs, the Kaplan-Meier curve is non parametric because we do not have assumption about distribution.

c.

```
ggsurvplot(km,fun="cumhaz",conf.int = F)
```



As we see probability of event increases while time pasts but not dramatically.

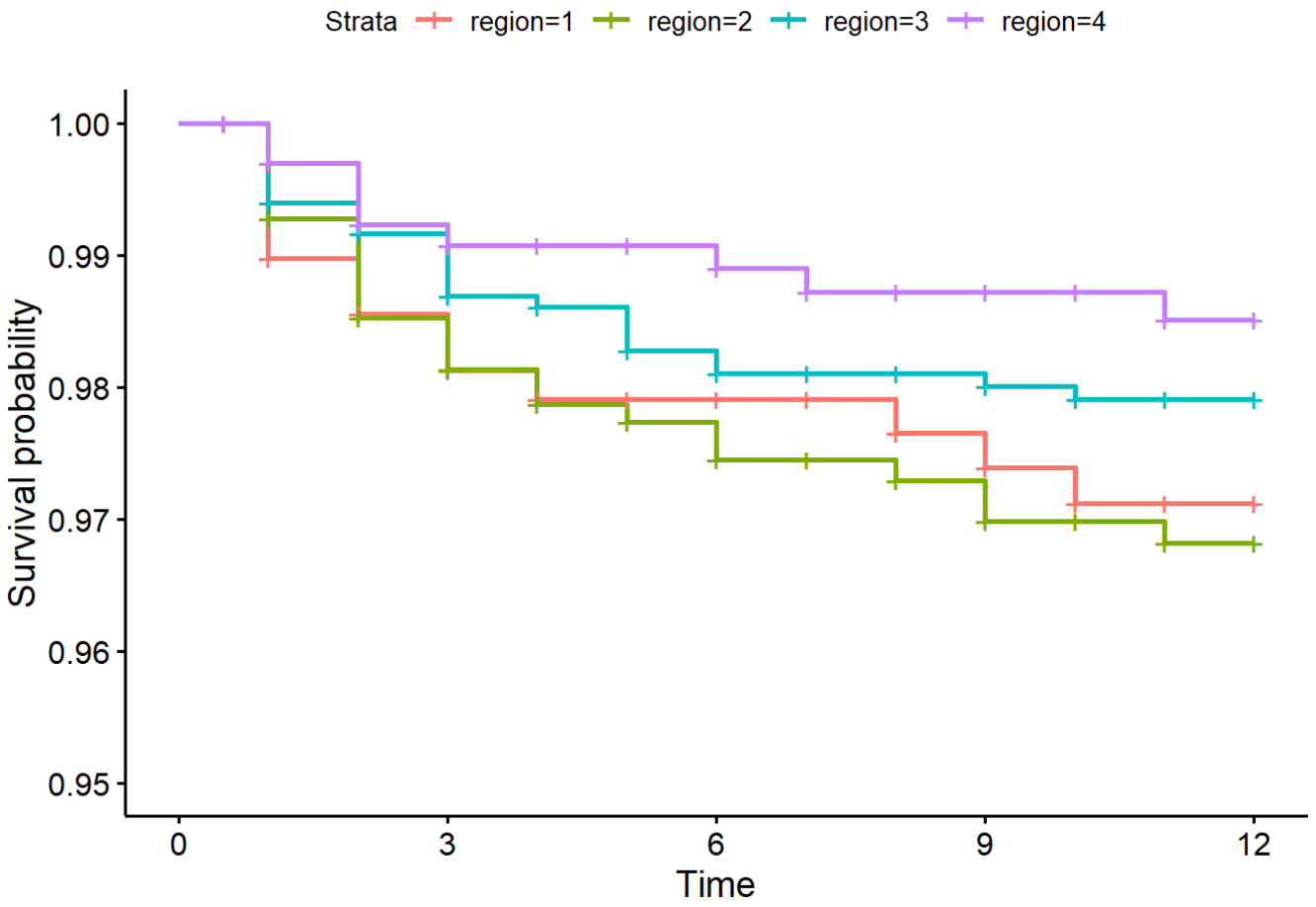
d.

```
str(df)
```

```
## 'data.frame':    3385 obs. of  14 variables:
##  $ chldage  : num  12 12 3 4 12 7 3 7 12 12 ...
##  $ hospital : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mthage   : int  22 20 24 21 20 24 24 26 24 27 ...
```

```
##  $ alcohol  : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 2 2 1 1 ...
##  $ smoke    : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 1 1 ...
##  $ region   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ poverty  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ bweight  : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 1 1 1 ...
##  $ race     : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ education: int  10 12 12 12 12 12 14 12 16 16 ...
##  $ nsibs    : int  1 1 2 0 0 1 0 0 0 1 ...
##  $ wmonth   : int  1 2 1 0 0 0 4 1 0 0 ...
##  $ sfmonth  : int  1 2 0 0 0 0 2 1 0 0 ...
##  $ agepn    : int  1 12 3 4 12 7 3 6 12 12 ...
##  - attr(*, "na.action")= 'omit' Named int  4 10 15 100 110 111 112 113 114 115 ...
##   ..- attr(*, "names")= chr  "4" "10" "15" "100" ...
```

```
km_region<-survfit(surv_obj~region,data=df)
ggsurvplot(km_region, conf.int = F,ylim=c(0.95,1))
```



e.

```
km_smoke<-survfit(surv_obj~smoke,data=df)
survdiff(surv_obj~smoke,data = df)
```

```
## Call:
## survdiff(formula = surv_obj ~ smoke, data = df)
##
```

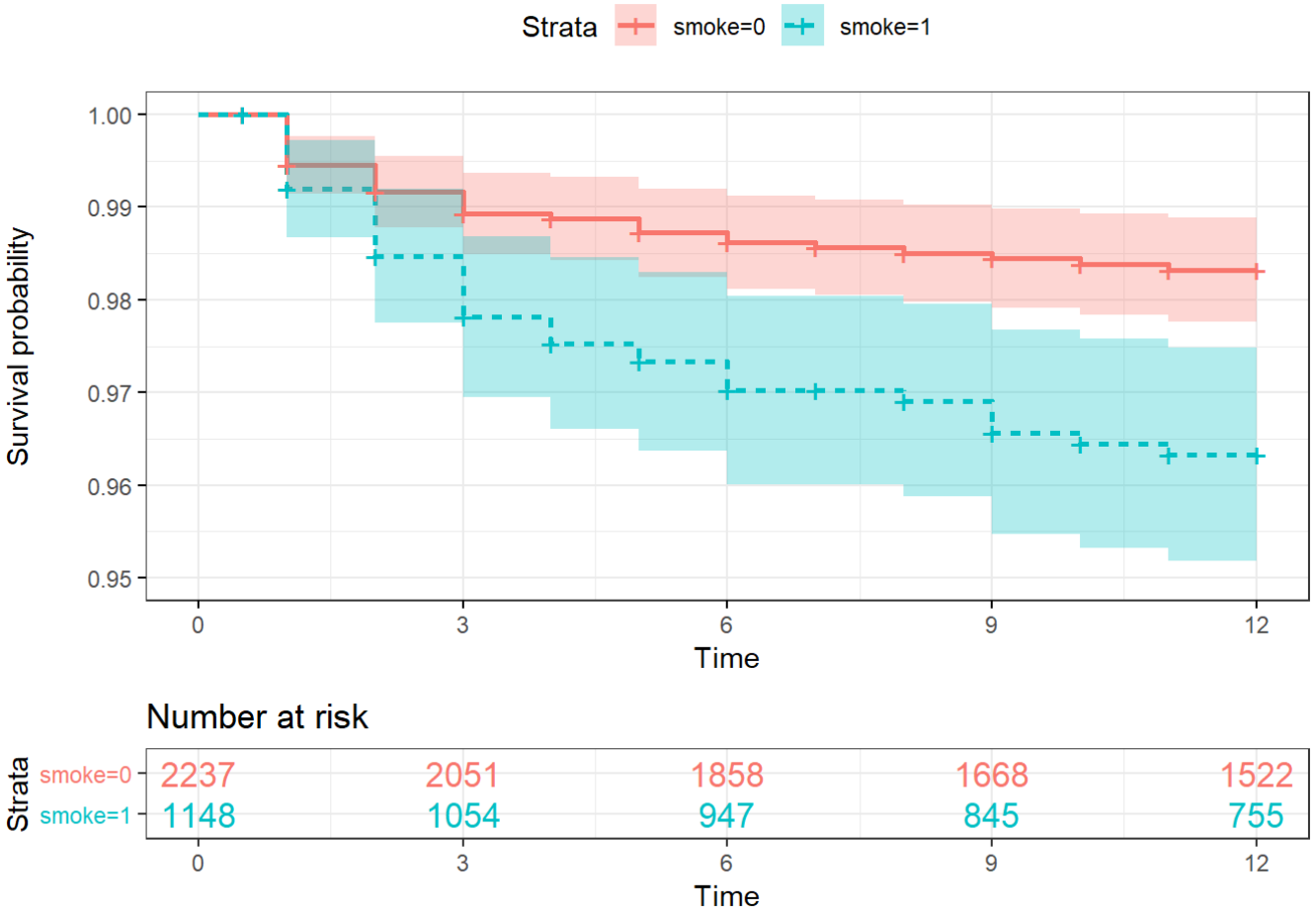
```
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## smoke=0 2237       34     47.6      3.89      11.5
## smoke=1 1148       38     24.4      7.58      11.5
##
##  Chisq= 11.5  on 1 degrees of freedom, p= 7e-04
```

```
ggsurvplot(km_smoke,ylim=c(0.95,1),
           pval = TRUE,
           conf.int = T,
           risk.table = TRUE, # Add risk table
           risk.table.col = "strata", # Change risk table color by groups
           linetype = "strata", # Change line type by groups
           surv.median.line = "hv", # Specify median survival, Allowed values include one of c("none", "hv", "h",
 "v"). v: vertical, h:horizontal.
           ggtheme = theme_bw())
```

```
## Warning in .add_surv_median(p, fit, type = surv.median.line, fun = fun, :
## Median survival not reached.
```

```
## Warning: Removed 1 rows containing missing values (geom_text).

## Warning: Removed 1 rows containing missing values (geom_text).
```



$H_0 : Survival\ curves\ are\ not\ different\ H_1 : Survival\ curves\ are\ different$ As $p_{value} < \alpha$ we reject H_0 and claim that Survival curves are different by smoke group. From the plot it is seen that when mother smokes probability of survival of child is decreasing compared with non smoking parent's child.

```
                                                Task 3   (4 pt.)
a. Run the ATF model with independent variables (at least one numeric and one categorical). Overall model and coe
fficients must be significant. Interpret the coefficients.
b. Make predictions based on your model using quantiles. Why do we not use the predict function?
c. Plot probability of a hazard and survival. Comment on it.
```

a.

```
str(df)
```

```
## 'data.frame':    3385 obs. of  14 variables:
##  $ chldage  : num  12 12 3 4 12 7 3 7 12 12 ...
##  $ hospital : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mthage   : int  22 20 24 21 20 24 24 26 24 27 ...
##  $ alcohol  : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 2 2 1 1 ...
##  $ smoke    : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 1 1 ...
##  $ region   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ poverty  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ bweight  : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 1 1 1 ...
##  $ race     : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ education: int  10 12 12 12 12 12 14 12 16 16 ...
##  $ nsibs    : int  1 1 2 0 0 1 0 0 0 1 ...
##  $ wmonth   : int  1 2 1 0 0 0 4 1 0 0 ...
##  $ sfmonth  : int  1 2 0 0 0 0 2 1 0 0 ...
##  $ agepn    : int  1 12 3 4 12 7 3 6 12 12 ...
##  - attr(*, "na.action")= 'omit' Named int  4 10 15 100 110 111 112 113 114 115 ...
##   ..- attr(*, "names")= chr  "4" "10" "15" "100" ...
```

```
unique(df$education)
```

```
##  [1] 10 12 14 16  9 11 13  8  1 15  7 17 19  6  0  3  4  2 18  5
```

```
ATF<-survreg(surv_obj~smoke+education,data = df,dist='exponential')
summary(ATF)
```

```
##
## Call:
## survreg(formula = surv_obj ~ smoke + education, data = df, dist = "exponential")
##               Value Std. Error     z       p
```

```
## (Intercept)  4.6195     0.6031  7.66 1.9e-14
## smoke1      -0.7068     0.2367 -2.99  0.0028
## education    0.1657     0.0538  3.08  0.0021
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -504   Loglik(intercept only)= -513.7
##  Chisq= 19.5 on 2 degrees of freedom, p= 5.8e-05
## Number of Newton-Raphson Iterations: 8
## n= 3385
```

```
exp(coef(ATF))
```

```
## (Intercept)      smoke1   education
## 101.4469144   0.4932265   1.1802016
```

The baseline is smoke 0(non smoking mother). So when a mother smokes the survival time is decreased by 50 percent. One unit increase in years of education increases survival time by 18 percent.

b.

```
pred<-predict(ATF,type='quantile',p=c(0.1,0.5,0.9))

head(pred)
```

```
##          [,1]     [,2]      [,3]
## [1,] 56.03749 368.6602 1224.6627
## [2,] 78.05326 513.4979 1705.8030
## [3,] 78.05326 513.4979 1705.8030
## [4,] 38.49794 253.2708  841.3473
## [5,] 78.05326 513.4979 1705.8030
## [6,] 78.05326 513.4979 1705.8030
```

There is 10% chance that survival is less than 56.03749 for first observation.And 50% chance that survival is less than 368.6602 and so on.

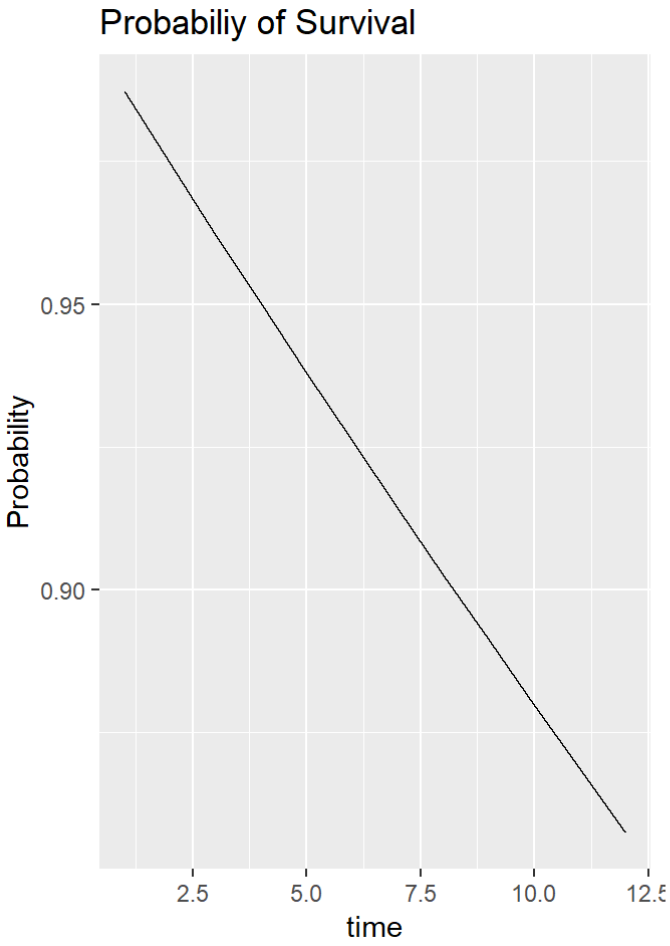
When we use predict(type=response) it provides us average of survival time. But it may not be that informative but when we know what type of distribution of survival time is we can calculate probability distribution or Cumulative probability distribution. And then we can find probabilities of survival given Time or find time given probabilities.

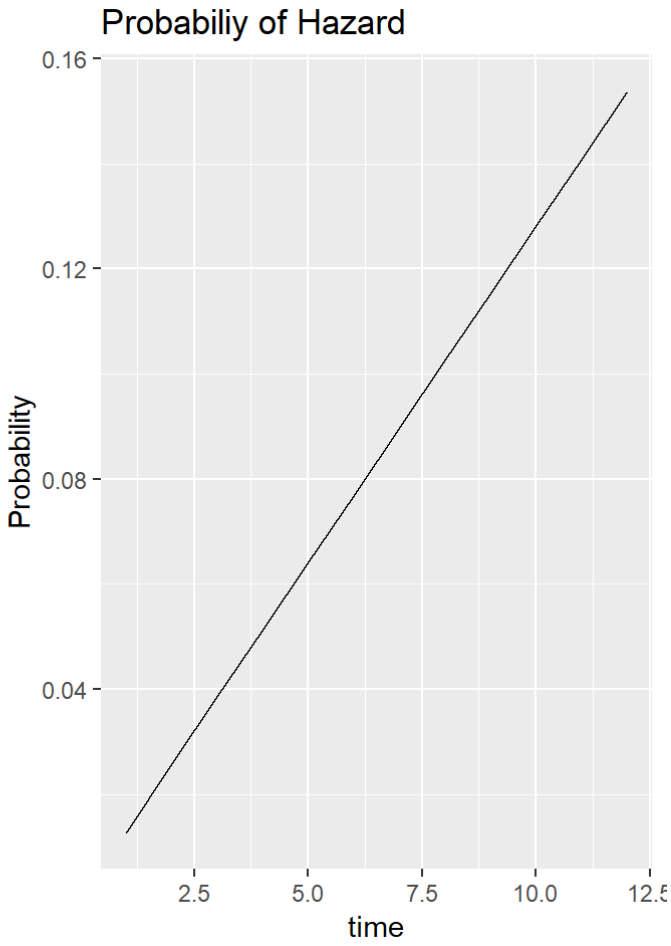
c. Plot probability of a hazard and survival. Comment on it.

```
library(gridExtra)
time<-1:12
pred_P<-pexp(time,rate=1/pred[2])
pred_P<-1-pred_P

g1<-ggplot()+geom_line(aes(x=time,y=pred_P))+
  labs(title = "Probabiliy of Survival",y="Probability")

g2<-ggplot()+geom_line(aes(x=time,y=-log(pred_P)))+
  labs(title = "Probabiliy of Hazard",y="Probability")
grid.arrange(g1, g2,nrow=1)
```





So while time passes probability of survival decreases and Hazard is increasing.

```
                                    Task 4   (6 pt.)

a. Which are the key assumptions of the Cox model?
b. Run the Cox proportional hazard model with significant coefficients. What does Rsquare show?
c. Interpret the coefficients of your model. What does HR indicate?
d. Check the proportional hazard assumption. Comment on it.
e. Make predictions with cox model using plots.
```

a. The basic Cox model assumes that the hazard functions for 2 different levels of a covariate are proportional for all values of time.
Proportional Hazard assumption:Hazard Ratio is independent from time.

b.

```
exp(0)
```

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

```
colnames(df)
```

```
##  [1] "chldage"   "hospital"  "mthage"    "alcohol"   "smoke"
##  [6] "region"    "poverty"   "bweight"   "race"      "education"
## [11] "nsibs"     "wmonth"    "sfmonth"   "agepn"
```

```
mod_cox<-coxph(surv_obj~smoke+education+wmonth,data=df)
summary(mod_cox)
```

```
## Call:
## coxph(formula = surv_obj ~ smoke + education + wmonth, data = df)
##
##   n= 3385, number of events= 72
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## smoke1     0.64567   1.90727  0.23686  2.726  0.00641 **
## education -0.14103   0.86846  0.05647 -2.498  0.01250 *
## wmonth    -0.22398   0.79933  0.08409 -2.664  0.00773 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## smoke1       1.9073     0.5243    1.1989    3.0341
## education    0.8685     1.1515    0.7775    0.9701
## wmonth       0.7993     1.2511    0.6779    0.9425
```

```
##
## Concordance= 0.697  (se = 0.029 )
## Rsquare= 0.01   (max possible= 0.289 )
## Likelihood ratio test= 32.55  on 3 df,   p=4e-07
## Wald test            = 24.19  on 3 df,   p=2e-05
## Score (logrank) test = 26.76  on 3 df,   p=7e-06
```

The Rsquare here is called a pseudo R square and shows the improvement of the model with predictors compared to the baseline model (no predictors). For pseudo Rsquare the upper limit is not necessary 1. The max possible Rsquare is 0.289 in this case and if our R square is close to this number it means the model is good. However, R square is 0.01 which indicates bad fit of model.

c.

```
exp(coef(mod_cox))
```

```
##    smoke1 education    wmonth
## 1.9072657 0.8684612 0.7993272
```

The hazard ratio (HR) is the ratio of the hazard rates corresponding to the conditions described by two levels of an explanatory variable of the same time. If we two observations where X2,X1 is bigger than by one unit the hazard ratio will be following

$$HR = \exp(\sum_{j=1}^{n} \beta_j (X_2 - X_1)$$

If difference between X2 and x1 is 1 the hazard ratio will be following. $HR = \exp(\beta)$ And if variable is not significant beta will be 0. So, $H_0: ()=1 $ $H_1: ()1 $ We reject H0 for all variables as p=value is less than alpha and can claim that all predictors in model are significant. Babies whose mother smokes has 90 percent more chance to be hospitalized compares those whose mother is non smoking. One unit increase in years of mother education decrease the hazard by 13%. One unit increase in Month the child was weaned decrease the hazard by 20%.

d.

```
test_z<-cox.zph(mod_cox,global = T,transform = 'rank')
test_z
```

```
##              rho chisq     p
## smoke1     0.103 0.752 0.386
## education  0.139 0.889 0.346
## wmonth    -0.086 0.633 0.426
## GLOBAL        NA 2.047 0.563
```

$H0 : Hazard\ rates\ are\ proportional\ H1 : Hazard\ rates\ are\ not\ proportional$ We fail to reject H0 for all variables so our hazard assumption are met.

e. Make predictions with cox model using plots.

```
str(df)
```

```
## 'data.frame':    3385 obs. of  14 variables:
##  $ chldage  : num  12 12 3 4 12 7 3 7 12 12 ...
##  $ hospital : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mthage   : int  22 20 24 21 20 24 24 26 24 27 ...
##  $ alcohol  : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 2 2 1 1 ...
##  $ smoke    : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 1 1 ...
##  $ region   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ poverty  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ bweight  : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 1 1 1 ...
##  $ race     : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ education: int  10 12 12 12 12 12 14 12 16 16 ...
##  $ nsibs    : int  1 1 2 0 0 1 0 0 0 1 ...
##  $ wmonth   : int  1 2 1 0 0 0 4 1 0 0 ...
##  $ sfmonth  : int  1 2 0 0 0 0 2 1 0 0 ...
##  $ agepn    : int  1 12 3 4 12 7 3 6 12 12 ...
##  - attr(*, "na.action")= 'omit' Named int  4 10 15 100 110 111 112 113 114 115 ...
##   ..- attr(*, "names")= chr  "4" "10" "15" "100" ...
```

```
case1<-data.frame(smoke="0",education=13,wmonth=3)
survivals<-survfit(mod_cox,newdata = case1)
df1<-data.frame(Probability=survivals$surv,Time=survivals$time)
df2<-data.frame(Probability=survivals$cumhaz,Time=survivals$time)
```

```
g1<-ggplot(df1,aes(Time,Probability))+geom_point()+ggtitle("Survival Probabilities")
g2<-ggplot(df2,aes(Time,Probability))+geom_point()+ggtitle("Cumulative Hazard Probabilities")
grid.arrange(g1, g2,nrow=1)
```