R instructions for the 11th seminar

Data set *Emamma.RData* contains data about 1000 female patients with breast cancer diagnosis treated at Masaryk Oncology Institute in Brno. The list of selected variables follows:
`AGE`: age when diagnosis was determined;
`TIME`: survival time in months;
`Death`: the status indicator, (0 - alive, 1 - dead);
`SIDE`: left or right;
`CHT`: chemotherapy (yes/no);
`CHT_Type`: type of chemotherapy (no chemotherapy, CMF, FAC, other);
`HT`: hormonal therapy (yes/no);
`LR`: local relapse (yes/no);
`MTS`: metastases (yes/no);
`MP`: menopause (0 - premenopausal, 1 - postmenopausal);
`HISTOL`: histology (1 - ductal, 2 - lobular, 3 - modular, 4 - other);
`STAGE`: stage of tumor disease (1, 2, 3, 4, higher values mean later stage)

```r
load("Emamma.RData") #Load the dataset first
library(survival) #Load the package needed for survival analysis
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```r
library(survminer)
```

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

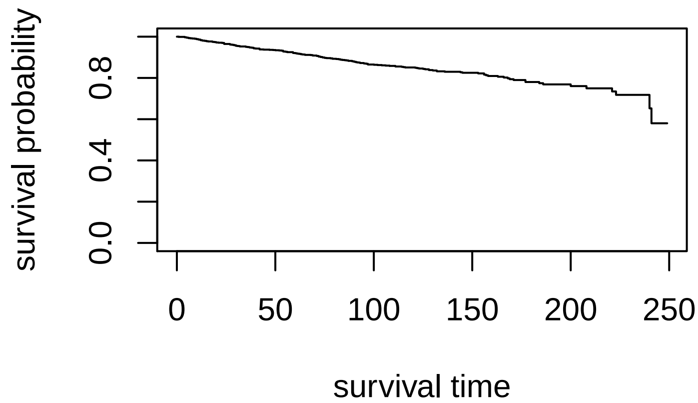
```
## Loading required package: magrittr
```

```r
library(ggfortify) #packages for better graphics
```

R Instructions for Problem 1:

1. Build the Kaplan-Meier estimate of the survival function for the whole dataset.

   ```r
   S <- Surv(Emamma$TIME, event = Emamma$Death)
   SResults <- survfit(S ~ 1, conf.type = "plain", type = "kaplan-meier")
   plot(SResults, conf.int = F, xlab = "survival time", ylab = "survival probability")
   ```



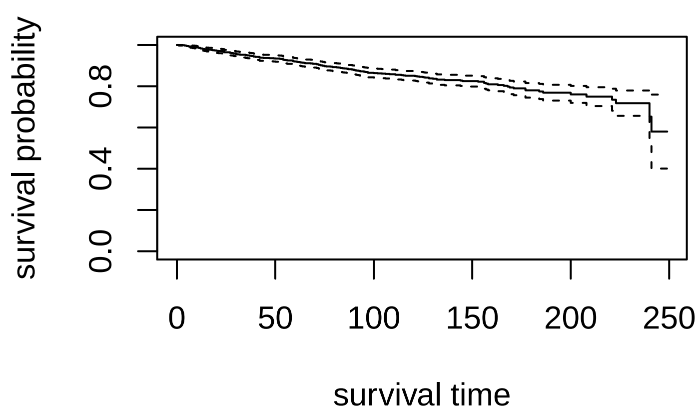
2. Find the median, lower and upper quartile for the survival time.

```
SResults
```

```
## Call: survfit(formula = S ~ 1, conf.type = "plain", type = "kaplan-meier")
##
##       n  events  median 0.95LCL 0.95UCL
##    1000     172      NA     241      NA
```

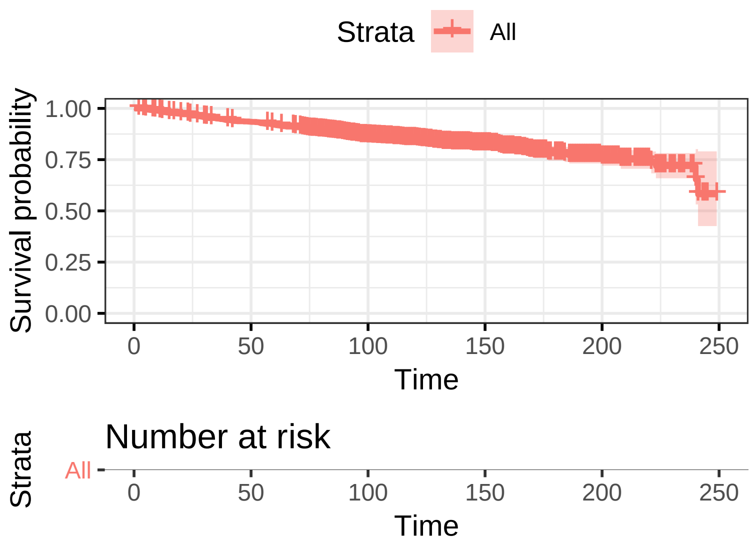
3. Create confidence intervals for the survival function. (Based on the formulae: $l = \hat{S}(t) - \sqrt{V\hat{a}r(\hat{S}(t))} \cdot u_{1-\alpha/2}$; $u = \hat{S}(t) + \sqrt{V\hat{a}r(\hat{S}(t))} \cdot u_{1-\alpha/2}$.) This can be done by setting the parameter `conf.int = TRUE`.

```
plot(SResults, conf.int = T, xlab = "survival time", ylab = "survival probability")
```



A better looking graph, with an optional risk table, can be created using the `ggsurvplot()` function, however, libraries `ggplot2` and `survminer` need to be installed first.

```r
ggsurvplot(survfit(S ~ 1), data = Emamma, ggtheme = theme_bw(), risk.table = T)
```



R Instructions for Problem 2:

1. Compare KM estimates of the survival function between groups of women in premenopausal state and in postmenopausal state. Which of these two groups is better off? Compare the median survival times for both groups.

```r
SResults_MP <- survfit(S ~ MP, data = Emamma)
SResults_MP
```

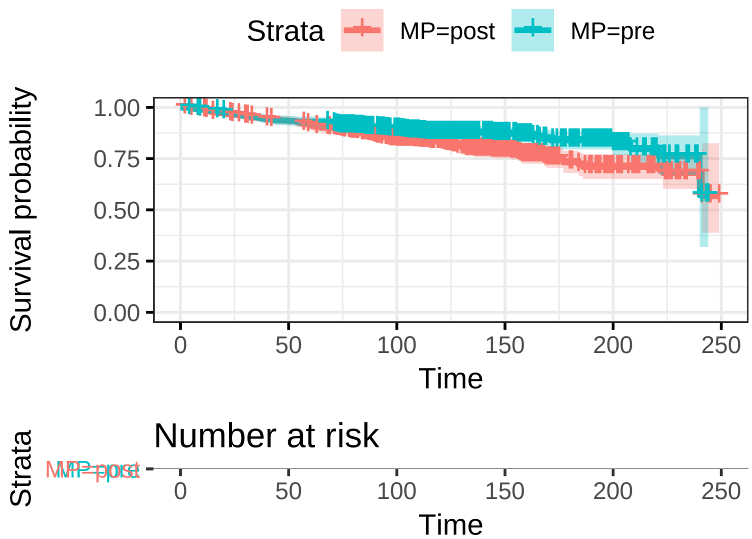
```
## Call: survfit(formula = S ~ MP, data = Emamma)
##
##            n events median 0.95LCL 0.95UCL
## MP=post 573    113     NA     241      NA
## MP=pre  427     59     NA     240      NA
```

```r
summary(SResults_MP, times = c(1, 30, 60, 90, 120, 150, 180, 210)) #manual choice of times
```

```
## Call: survfit(formula = S ~ MP, data = Emamma)
##
##                 MP=post
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    573       0    1.000 0.00000        1.000        1.000
##    30    539      25    0.956 0.00862        0.939        0.973
##    60    506      25    0.911 0.01200        0.888        0.935
##    90    426      26    0.862 0.01476        0.833        0.891
##   120    277      13    0.832 0.01643        0.800        0.865
##   150    155      12    0.790 0.01969        0.752        0.829
##   180     74       8    0.731 0.02741        0.680        0.787
##   210     37       2    0.709 0.03090        0.651        0.772
##
##                 MP=pre
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    427       1    0.998 0.00234        0.993        1.000
##    30    405      18    0.955 0.01006        0.936        0.975
```

```
##     60     393       10     0.931 0.01229          0.908          0.956
##     90     338       13     0.900 0.01468          0.871          0.929
##    120     225        8     0.876 0.01659          0.844          0.909
##    150     151        1     0.870 0.01735          0.837          0.905
##    180      83        4     0.839 0.02290          0.795          0.885
##    210      30        2     0.795 0.03743          0.725          0.872
```

```r
ggsurvplot(survfit(S ~ MP, data = Emamma), conf.int = TRUE, ggtheme = theme_bw(),
           risk.table = T)
```



From these results, mainly from the number of events for various times and from plotting the categorized survival function, we can see that pre-menopausal women have a higher chance of survival than women in post-menopausal state.
Explanation of the `NA` medians - If one of the groups has not yet dropped to 50% survival at the end of the available data, we cannot compute a median survival and there will be `NA` values for median survival produced in such cases. In our case, neither of the two groups dropped to 50% survival, which can also be seen in the plots - therefore the `NA` medians.

2. Use the log-rank test to test the null hypothesis that the survival functions for these two groups are the same.

```r
log_rank_MP <- survdiff(S ~ MP, data = Emamma)
log_rank_MP
```

```
## Call:
## survdiff(formula = S ~ MP, data = Emamma)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## MP=post 573      113     95.8       3.1      7.03
## MP=pre  427       59     76.2       3.9      7.03
##
##  Chisq= 7  on 1 degrees of freedom, p= 0.008
```

The resulting p-value of this test is p = 0,008, which means that we reject the null hypothesis that the survival functions for women with different menopausal states are equal (we consider 95% level of significance throughout the whole analysis).

3. Compare KM estimates of the survival function between groups with different types of chemotherapy.

Use the log-rank test to test the null hypothesis that the survival functions for these groups are the same (χ^2 statistic now has 3 degrees of freedom = the number of levels of the qualitative variable that we categorize by, minus 1).

```r
SResults_CHT <- survfit(S ~ CHT_Type, data = Emamma)
SResults_CHT
```

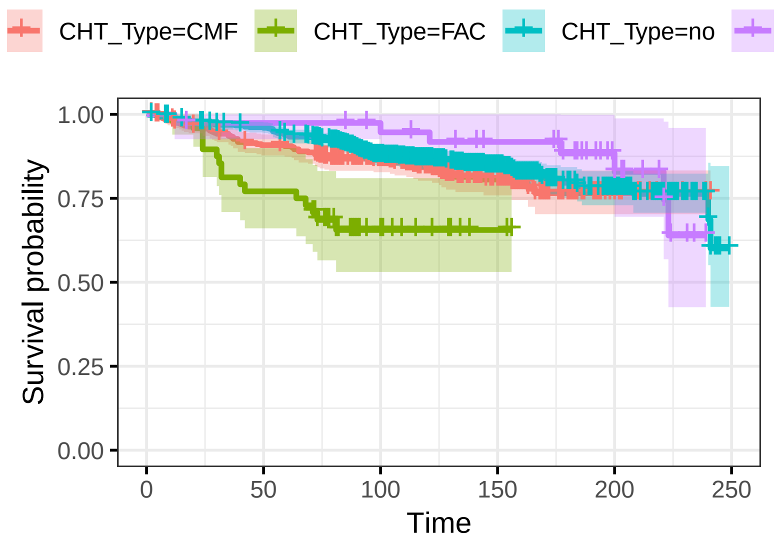
```
## Call: survfit(formula = S ~ CHT_Type, data = Emamma)
##
##                   n events median 0.95LCL 0.95UCL
## CHT_Type=CMF    333     58     NA      NA      NA
## CHT_Type=FAC     48     16     NA      NA      NA
## CHT_Type=no     580     91     NA     241      NA
## CHT_Type=other   39      7     NA     223      NA
```

```r
summary(SResults_CHT, times = c(1, 30, 60, 90, 120, 150, 180, 210))
```

```
## Call: survfit(formula = S ~ CHT_Type, data = Emamma)
##
##                 CHT_Type=CMF
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    333       0    1.000  0.0000        1.000        1.000
##    30    309      19    0.942  0.0129        0.917        0.968
##    60    293      12    0.905  0.0162        0.874        0.938
##    90    252      12    0.868  0.0188        0.832        0.906
##   120    164       6    0.842  0.0209        0.802        0.884
##   150     96       6    0.806  0.0248        0.759        0.856
##   180     33       3    0.765  0.0332        0.703        0.833
##   210      8       0    0.765  0.0332        0.703        0.833
##
##                 CHT_Type=FAC
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     48       0    1.000  0.0000        1.000        1.000
##    30     43       6    0.875  0.0477        0.786        0.974
##    60     37       5    0.771  0.0607        0.661        0.899
##    90     16       5    0.656  0.0707        0.531        0.810
##   120      7       0    0.656  0.0707        0.531        0.810
##   150      2       0    0.656  0.0707        0.531        0.810
##
##                 CHT_Type=no
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    580       1    0.998 0.00172        0.995        1.000
##    30    555      17    0.969 0.00725        0.955        0.983
##    60    532      18    0.937 0.01015        0.917        0.957
##    90    460      22    0.896 0.01294        0.871        0.922
##   120    298      14    0.866 0.01477        0.838        0.896
##   150    179       6    0.845 0.01682        0.813        0.879
##   180    100       8    0.796 0.02322        0.752        0.843
##   210     47       3    0.763 0.02966        0.707        0.823
##
##                 CHT_Type=other
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     39       0    1.000  0.0000        1.000        1.000
##    30     37       1    0.974  0.0253        0.926        1.000
##    60     37       0    0.974  0.0253        0.926        1.000
```

```
##     90      36        0     0.974  0.0253         0.926         1.000
##    120      33        1     0.947  0.0368         0.877         1.000
##    150      29        1     0.918  0.0455         0.833         1.000
##    180      24        1     0.884  0.0551         0.782         0.999
##    210      12        1     0.829  0.0744         0.695         0.988
```

```
ggsurvplot(survfit(S ~ CHT_Type, data= Emamma), conf.int= TRUE, ggtheme=theme_bw())
```




```
log_rank_CHT <- survdiff(S ~ CHT_Type, data = Emamma)
log_rank_CHT
```

```
## Call:
## survdiff(formula = S ~ CHT_Type, data = Emamma)
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## CHT_Type=CMF     333       58    53.64     0.354     0.521
## CHT_Type=FAC      48       16     5.34    21.276    22.185
## CHT_Type=no      580       91   103.50     1.510     3.821
## CHT_Type=other    39        7     9.52     0.666     0.722
##
##  Chisq= 24.2  on 3 degrees of freedom, p= 2e-05
```

As we can see from the results of the second log-rank test, the null hypothesis stating that the survival functions for different types of chemoteraphy are equal, should also be rejected. Previous graph also suggests significant differences.

R Instructions for Problem 3:

Create a Cox model, where survival time depends on the variables `AGE`, `CHT` and `MP`.

Here we have to adjust one observation with a "missing" value ("") in the column `CHT`, which is treated as another level and due to the fact that `R` automatically sets this "missing" value as a reference level for the variable `CHT`. This could cause misleading results. We can resolve this by setting the empty string to `NA` instead and then dropping the unused factor levels with the function `droplevels()`.

```
levels(Emamma$CHT)
```

```
## [1] ""    "no"  "yes"
```

```r
which(Emamma$CHT == "")
```

```
## [1] 471
```

```r
Emamma[471, "CHT"] <- NA
Emamma$CHT <- droplevels(Emamma$CHT)

cox <- coxph(S ~ Age + CHT + MP, data = Emamma, ties = "efron")
summary(cox)
```

```
## Call:
## coxph(formula = S ~ Age + CHT + MP, data = Emamma, ties = "efron")
##
##   n= 999, number of events= 172
##    (1 observation deleted due to missingness)
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## Age    0.05849   1.06023  0.01147 5.098 3.43e-07 ***
## CHTyes 0.66940   1.95306  0.16980 3.942 8.07e-05 ***
## MPpre  0.32594   1.38533  0.23983 1.359    0.174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## Age        1.060     0.9432    1.0367     1.084
## CHTyes     1.953     0.5120    1.4002     2.724
## MPpre      1.385     0.7218    0.8658     2.217
##
## Concordance= 0.633  (se = 0.022 )
## Likelihood ratio test= 40.63  on 3 df,   p=8e-09
## Wald test            = 38.76  on 3 df,   p=2e-08
## Score (logrank) test = 39.18  on 3 df,   p=2e-08
```

1. Test the statistical significance of a full model by the means of likelihood ratio test ($H_0 : \beta_0 = \dots = \beta_k = 0$).

   The output of `summary(cox)` gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score logrank statistics, which are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ a bit. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred. Here we can see that all of the p-values show statistical significance.

2. Estimate and interpret the parameters β. Estimate the relative risk.

   Values of the estimated parameters are displayed in the column `coef`. A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable.

   ```r
   cox$coefficients
   ```

   ```
   ##        Age     CHTyes      MPpre
   ## 0.05848903 0.66939746 0.32593913
   ```

3. Which of the parameters are statistically non-significant, according to Wald's test statistic? Based on this result, which of the variables can be left out of the model? (MP)

   The column marked "`z`" gives the Wald statistic value, which evaluates, whether the β coefficient of a given variable is statistically significantly different from 0. Here we can see that only the variables `Age`

and `CHT` seem to be statistically significant.
The exponentiated coefficients (`exp(coef)`), also known as hazard ratios, give the effect size of covariates.

4. Test the significance of the variable MP using likelihood ratio test.

   Here we first have to load the library `lmtest` and then we can use the `lrtest()` function to perform likelihood ratio test on two nested models.

```
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

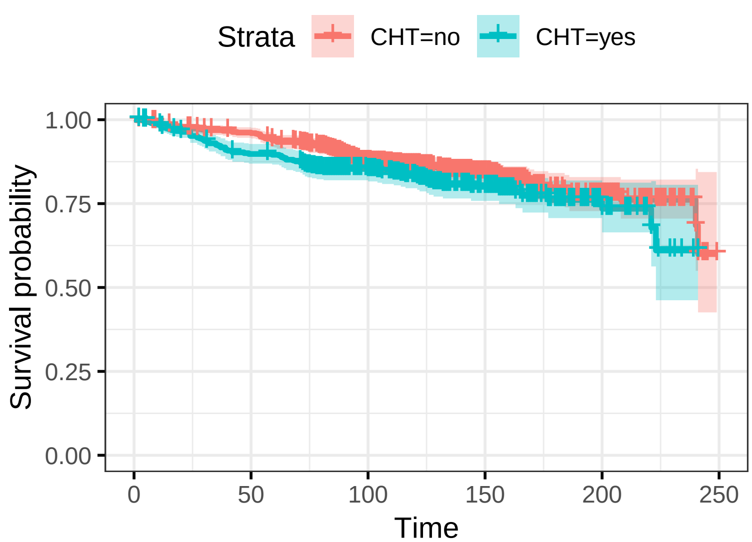
```
cox_2 <- coxph(Surv(TIME, event = Death) ~ Age + CHT, data = Emamma, ties = "efron")
lrtest(cox, cox_2)
```

```
## Likelihood ratio test
##
## Model 1: S ~ Age + CHT + MP
## Model 2: Surv(TIME, event = Death) ~ Age + CHT
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -1094.5
## 2   2 -1095.4 -1 1.8529     0.1735
```

   From the results of the likelihood ratio test, we can see that the variable `MP` indeed seems to be insignificant.

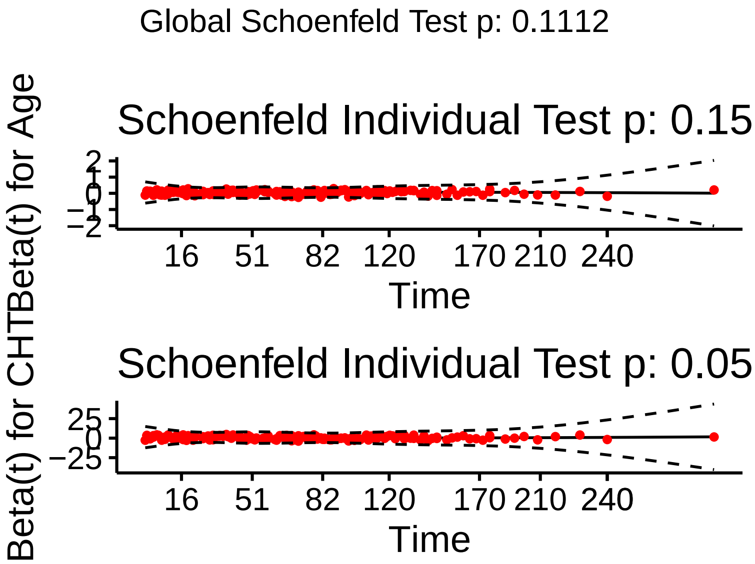
5. Leave the variable MP out of the model. Calculate the average age (when the diagnosis was determined) and plot the survival function based on your estimates of parameters β, with the following values for regressors: `AGE` = average of the whole sample, `CHT` = 0 and `CHT` = 1.

```
age_avg <- mean(Emamma$Age)
ggsurvplot(survfit(S ~ CHT, data = Emamma), conf.int = T, ggtheme = theme_bw())
```



6. Using scaled Schoenfeld's residuals, first by the variable `AGE` and then by the variable `CHT` determine whether the assumption of proportional hazards (the hazard ratio should be constant over time) was fulfiled. (In a graph, Schoenfeld's residuals should be on the y-axis, $\ln(t)$ on the x-axis and an intersected line should be identical to the x-axis.)

```
test.ph <- cox.zph(cox_2)
ggcoxzph(test.ph)
```



Based on these results, we can assume that the hazard ratio is constant over time, which means that the necessary assumption of proportional hazards is fulfilled.

R Instructions for Problem 4:

Create a Cox model, where survival time depends on the variables `AGE`, `CHT`, `MP` and `HISTOL`. For the variable `HISTOL`, which has 4 levels, set `ductal` as a reference level.

Repeat the same steps as in Problem 3

1. Test the statistical significance of a full model by the means of likelihood ratio test ($H_0 : \beta_0 = \cdots = \beta_k = 0$).
2. Which of the parameters are statistically non-significant, according to Wald's test statistic? Based on this result, which of the variables can be left out of the model?
3. Interpret the parameters β and the relative risk. (There will be three "betas" for the categorical variable `HISTOL`, which will be interpreted as: "How will the risk change, when we change the histology from ductal to lobular/modular?,…")

Voluntary: Adding the variable `STAGE` is even more interesting. (You will see that with higher stage, the risk increases rapidly.)

Note to regression models: An epidemiologist, who examines the whole population, might be only interested in knowing the survival time. However, a doctor treating a specific patient wants to know the survival time according to specific values of the regressors of his patient.