R code for Survival Lecture #1

Spring 2025

Contents

Introduction

This document demonstrates how to estimate and visualize Kaplan–Meier survival curves using R. We will first compute an overall survival estimate, then compare survival between treatment arms ("Placebo" vs. "6-MP") in the `leuk` dataset.

To install a package in R, you can follow these steps:

1. Open R or RStudio.
2. Click on the "Packages" tab in the bottom right pane, or type library() in the console to see a list of installed packages.
3. If the package you want to install is not already installed, type install.packages("package_name") in the console, replacing "package_name" with the name of the package you want to install.
4. Press enter, and R will start downloading and installing the package and any dependencies. This may take a few minutes depending on the size of the package and your internet speed.
5. Once the package is installed, you can load it into your workspace using the library() function, like this: library(package_name).

Load and Inspect Data

```
library(haven)    # read Stata (.dta) files

leuk <- read_dta("leuk.dta")
head(leuk)          # view variables in the dataset
```

```
## # A tibble: 6 x 4
##      id  time  cens group
##   <dbl> <dbl> <dbl> <dbl+lbl>
## 1    28     6     1 1 [6-MP]
## 2    34     6     1 1 [6-MP]
## 3    40     6     0 1 [6-MP]
## 4    12     6     1 1 [6-MP]
## 5     4     7     1 1 [6-MP]
## 6    38     9     0 1 [6-MP]
```

```
summary(leuk)     # basic descriptive statistics
```

```
##        id             time            cens            group
##  Min.   : 1.00   Min.   : 1.00   Min.   :0.0000   Min.   :0.0
##  1st Qu.:11.25   1st Qu.: 6.00   1st Qu.:0.0000   1st Qu.:0.0
##  Median :21.50   Median :10.50   Median :1.0000   Median :0.5
##  Mean   :21.50   Mean   :12.88   Mean   :0.7143   Mean   :0.5
##  3rd Qu.:31.75   3rd Qu.:18.50   3rd Qu.:1.0000   3rd Qu.:1.0
##  Max.   :42.00   Max.   :35.00   Max.   :1.0000   Max.   :1.0
```

Kaplan-Meier Estimate (6-MP group only)

The survival package is then loaded to perform survival analyses. The survfit() function is used to fit a Kaplan-Meier estimator to the survival data in the 6-MP group with the Surv() function specifying the time and censoring variables. The resulting Kaplan-Meier estimator is stored as km_6mp, and the summary() function is used to print a summary of the estimator.

```
library(survival)    # load core survival analysis functions
library(survminer)   # load high-level plotting tools for survival objects

# Convert the numeric treatment code into a factor with descriptive labels
leuk$group <- factor(leuk$group,
                     levels = c(0, 1),
                     labels = c("Placebo", "6-MP"))

# Kaplan-Meier estimate restricted to the "6-MP" treatment arm only
km_6mp <- survfit(Surv(time, cens) ~ 1,
                  data = subset(leuk, group == "6-MP"))

# Print detailed survival estimates for the 6-MP group:
# time = follow-up times
# n.risk = patients still at risk at each time
# n.event = events (relapses/deaths) at each time
# surv = estimated survival probability
# lower/upper = 95% confidence limits
summary(km_6mp)
```

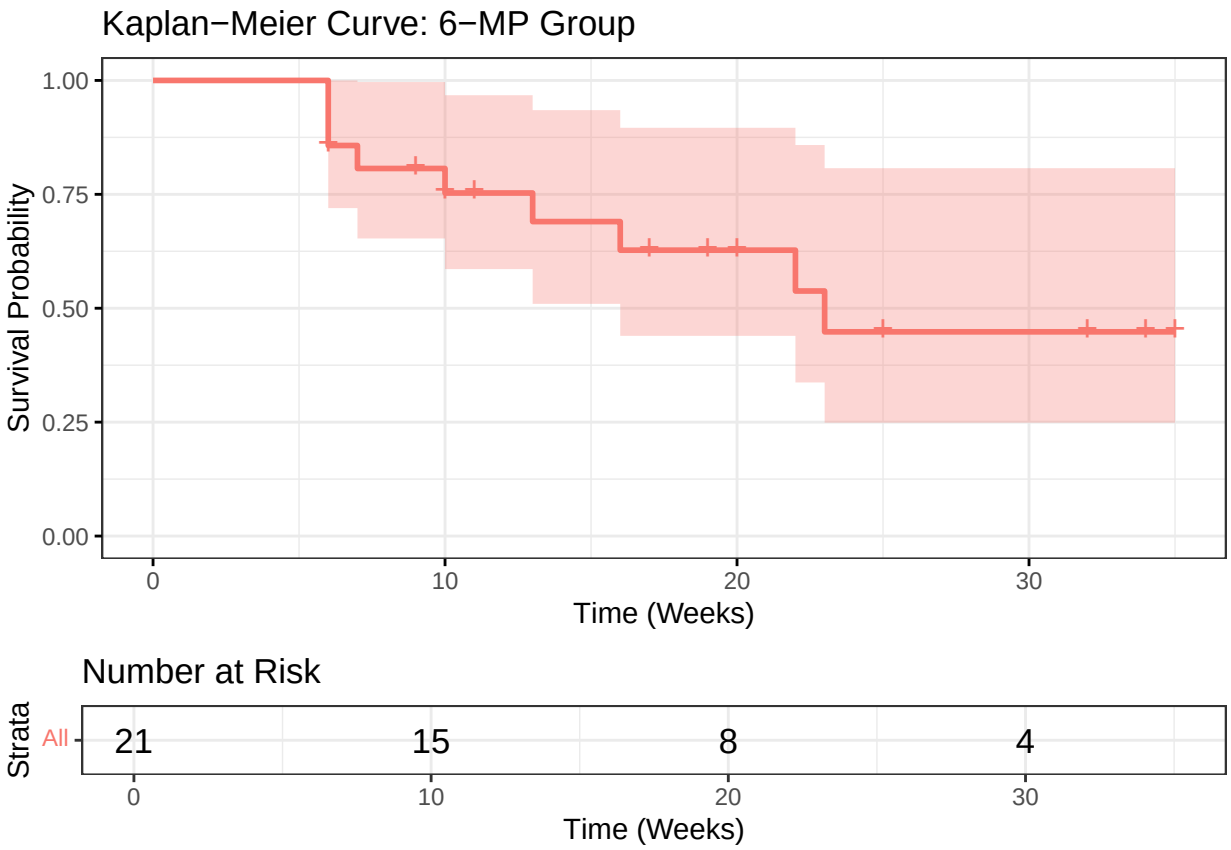
```
## Call: survfit(formula = Surv(time, cens) ~ 1, data = subset(leuk, group ==
##     "6-MP"))
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.720        1.000
##     7     17       1    0.807  0.0869        0.653        0.996
##    10     15       1    0.753  0.0963        0.586        0.968
##    13     12       1    0.690  0.1068        0.510        0.935
##    16     11       1    0.627  0.1141        0.439        0.896
##    22      7       1    0.538  0.1282        0.337        0.858
##    23      6       1    0.448  0.1346        0.249        0.807
```

```
# median survival time (the time at which estimated survival = 0.5)
# along with its 95% confidence interval for the 6-MP group
km_6mp
```

```
## Call: survfit(formula = Surv(time, cens) ~ 1, data = subset(leuk, group ==
##     "6-MP"))
##
##       n events median 0.95LCL 0.95UCL
## [1,] 21      9     23      16      NA
```

```
# Kaplan-Meier survival curve for the "6-MP" treatment group only
ggsurvplot(km_6mp,
           data = leuk,
```

```
            censor = TRUE,
            xlab = "Time (Weeks)",
            ylab = "Survival Probability",
            title = "Kaplan-Meier Curve: 6-MP Group",
            break.time.by = 10,
            legend = "none",              # remove legend (only one curve)
            risk.table = TRUE,             # display number at risk below the plot
            risk.table.title = "Number at Risk",
            risk.table.y.text.col = TRUE,
            risk.table.height = 0.25,     # allocate 25% of total figure height to the risk table
            ggtheme = theme_bw()
)
```



Kaplan-Meier Estimate by Treatment Group

```r
# Fit Kaplan-Meier survival curves stratified by treatment group
km_group <- survfit(Surv(time, cens) ~ group, data = leuk, conf.type = "log-log")

# Print summary of stratified survival fit
summary(km_group)
```

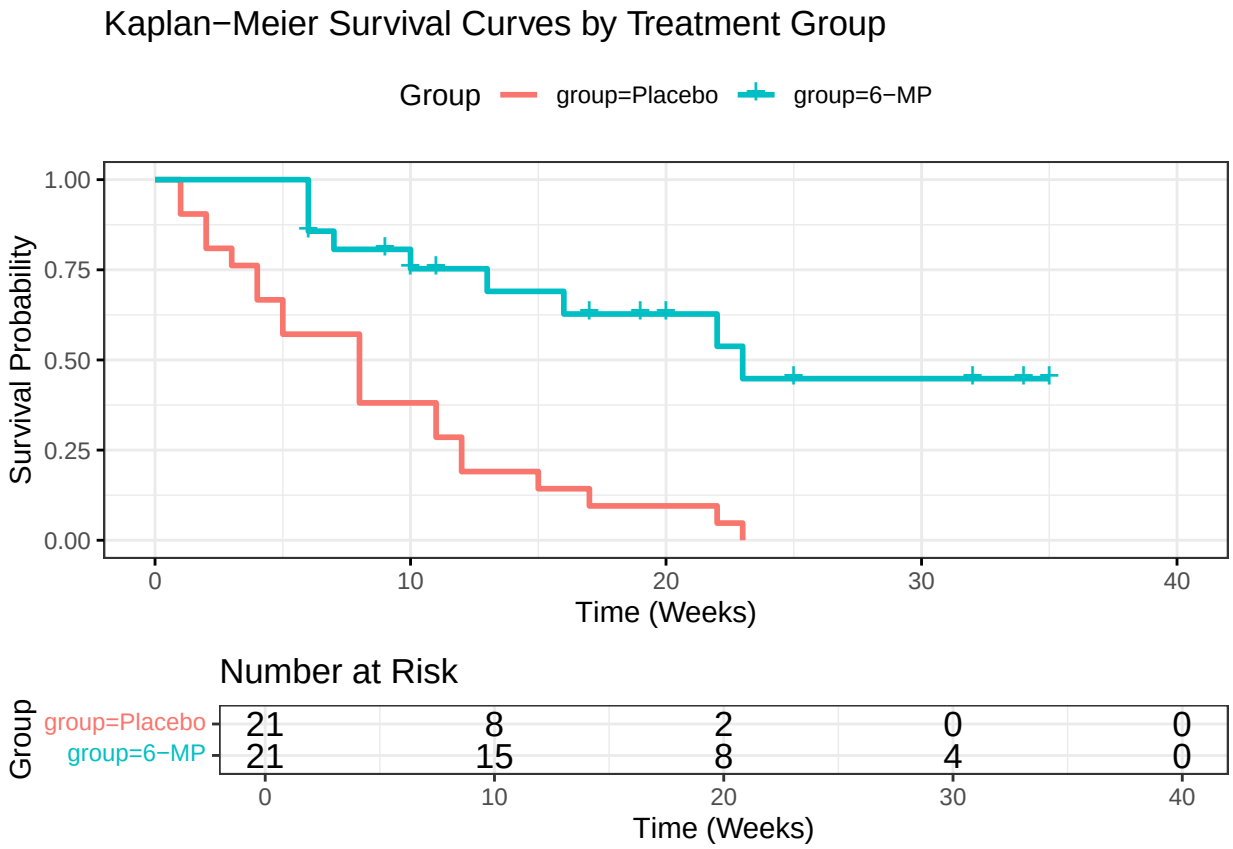
```
## Call: survfit(formula = Surv(time, cens) ~ group, data = leuk, conf.type = "log-log")
##
##                 group=Placebo
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     21       2   0.9048  0.0641      0.67005        0.975
##     2     19       2   0.8095  0.0857      0.56891        0.924
##     3     17       1   0.7619  0.0929      0.51939        0.893
##     4     16       2   0.6667  0.1029      0.42535        0.825
##     5     14       2   0.5714  0.1080      0.33798        0.749
##     8     12       4   0.3810  0.1060      0.18307        0.578
##    11      8       2   0.2857  0.0986      0.11656        0.482
##    12      6       2   0.1905  0.0857      0.05948        0.377
##    15      4       1   0.1429  0.0764      0.03566        0.321
##    17      3       1   0.0952  0.0641      0.01626        0.261
##    22      2       1   0.0476  0.0465      0.00332        0.197
##    23      1       1   0.0000     NaN           NA           NA
##
##                 group=6-MP
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.620        0.952
##     7     17       1    0.807  0.0869        0.563        0.923
##    10     15       1    0.753  0.0963        0.503        0.889
##    13     12       1    0.690  0.1068        0.432        0.849
##    16     11       1    0.627  0.1141        0.368        0.805
##    22      7       1    0.538  0.1282        0.268        0.747
##    23      6       1    0.448  0.1346        0.188        0.680
```

```r
# Median survival time (with 95% CI) for each treatment group
km_group
```

```
## Call: survfit(formula = Surv(time, cens) ~ group, data = leuk, conf.type = "log-log")
##
##                n events median 0.95LCL 0.95UCL
## group=Placebo 21     21      8       4      11
## group=6-MP    21      9     23      13      NA
```

```r
# Output columns:
# strata - treatment group
# median - estimated median survival time (time when survival = 0.5)
# 0.95LCL & 0.95UCL - lower and upper bounds of the 95% confidence interval
```

```r
# Plot stratified Kaplan-Meier curves
ggsurvplot(km_group,
           data = leuk,
           censor = TRUE,
           xlab = "Time (Weeks)",
           ylab = "Survival Probability",
           title = "Kaplan-Meier Survival Curves by Treatment Group",
           legend.title = "Group",
           break.time.by = 10,
           risk.table = TRUE,           # display number at risk below the plot
           risk.table.title = "Number at Risk",
           risk.table.y.text.col = TRUE,
           risk.table.height = 0.25,    # allocate 25% of total figure height to the risk table
           ggtheme = theme_bw())
```



Log-rank test

The survdiff() function is used to perform a log-rank test by group, comparing the survival curves of the two groups.

```
logrank_test <- survdiff(Surv(time, cens) ~ group, data = leuk)

# Print the test results:
#  chisq = test statistic
#  p-value (calculated from chi-square) indicates whether survival differs by group
logrank_test
```

```
## Call:
## survdiff(formula = Surv(time, cens) ~ group, data = leuk)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## group=Placebo 21       21     10.7      9.77      16.8
## group=6-MP    21        9     19.3      5.46      16.8
##
##  Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

Wilcoxon test — gives more weight to early events

```
wilcoxon_test <- survdiff(Surv(time, cens) ~ group,
                          data = leuk,
                          rho = 1)
# Print the test results:
#  chisq = test statistic
#  p-value (calculated from chi-square) indicates whether survival differs by group
wilcoxon_test
```

```
## Call:
## survdiff(formula = Surv(time, cens) ~ group, data = leuk, rho = 1)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## group=Placebo 21    14.55     7.68      6.16      14.5
## group=6-MP    21     5.12    12.00      3.94      14.5
##
##  Chisq= 14.5  on 1 degrees of freedom, p= 1e-04
```

Cox regression

The coxph() function is used to fit a Cox proportional hazards model with group as the only predictor variable, and the resulting object is passed into the tbl_regression() function from the gtsummary package to create a summary table of the Cox model output with exponentiated coefficients and 95% confidence intervals.

```
# Fit a Cox proportional hazards model with treatment group as predictor
cox_fit <- coxph(Surv(time, cens) ~ group, data = leuk)

# Print model summary:
# coef = log(hazard ratio) comparing 6-MP vs Placebo
# exp(coef) = hazard ratio (HR)
# se(coef) = standard error of the coefficient
# z = Wald test statistic
# Pr(>|z|) = p-value for testing HR = 1
# Additionally, the summary provides 95% CI for HR,
# a concordance statistic reflecting the model's predictive accuracy,
# and results from global tests (Likelihood ratio, Wald, and Score tests)
# for assessing the overall covariate effect

summary(cox_fit)
```

```
## Call:
## coxph(formula = Surv(time, cens) ~ group, data = leuk)
##
##   n= 42, number of events= 30
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## group6-MP -1.5721    0.2076   0.4124 -3.812 0.000138 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## group6-MP    0.2076      4.817   0.09251    0.4659
##
## Concordance= 0.69  (se = 0.041 )
## Likelihood ratio test= 16.35  on 1 df,   p=5e-05
## Wald test            = 14.53  on 1 df,   p=1e-04
## Score (logrank) test = 17.25  on 1 df,   p=3e-05
```

```
# Present results in a clean table with exponentiated coefficients (HRs) and 95% CIs
library(gtsummary)
tbl_regression(cox_fit,
               exponentiate = TRUE,    # show hazard ratios instead of log(HR)
               conf.level = 0.95)      # 95% confidence intervals
```

Characteristic	HR	95% CI	p-value
group			
Placebo	—	—	
6-MP	0.21	0.09, 0.47	<0.001