Introduction to Survival Analysis

F.A. BarriosInstituto de Neurobiología UNAM

2021-01-11

```
library(tidyverse)
library(survival)
library(survminer)
library(car)

setwd("~/Dropbox/GitHub/Class2020")
leuk <- read_csv(file="DataRegressBook/Chap3/leuk.csv")
# wcgs <- read_csv("DataRegressBook/Chap2/wcgs.csv")
```

Introduction to Survival Analysis

Survival analysis is used to analyze data in which the time until the event is of interest. The response variable is the time until that event and is often called a *failure time*, *survival time*, or *event time*. The response, event time, is usually continuous, but survival analysis allows the response to be incompletely determined for some subjects. Then we say that the survival time is *censored* on the right.

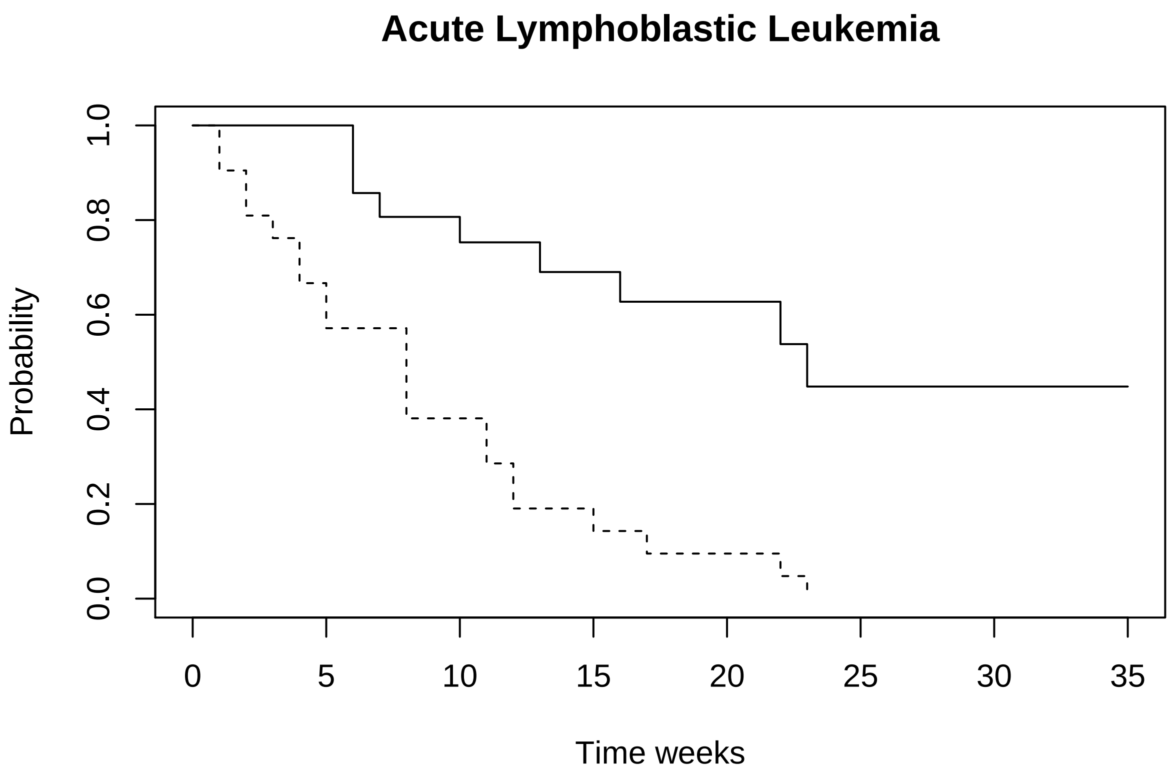
There are several reasons for studying failure time using the specialized methods of survival analysis.
1. Time to failure can have an unusual distribution. Failure time is restricted to be positive so it has a skewed distribution and will never be normally distributed.
2. The probability of surviving past a certain time is often more relevant than the expected survival time (and expected survival time may be difficult to estimate if the amount of censoring is large).
3. A function used in survival analysis, the hazard function, helps one to understand the mechanism of failure

Right Censoring

To illustrate the special characteristics of survival data, we consider a study of *6-mercaptopurine* (6-MP) as maintenance therapy for children in remission from *acute lymphoblastic leukemia* (ALL) Forty-two patients achieved remission from induction therapy and were then randomized in equal numbers to 6-MP or placebo. The survival time studied was from randomization until relapse. At the time of the analysis, all 21 patients in the placebo group had relapsed, whereas only 9 of 21 patients in the 6-MP group had. One crucial characteristic of these survival times is that for the 12 patients in the 6-MP group who remained in remission at the time of the analysis, the exact time to relapse was unobserved; it was only known to exceed the follow-up time. For example, one patient had only been under observation for six weeks, so we only know that the relapse time is longer than that. Such a survival time is said to be *right-censored*.

Definition: A survival time is said to be right-censored at time t if it is only known to be greater than t.

```
plot(survfit(Surv(time, cens) ~ group, data = leuk), main = "Acute Lymphoblastic Leukemia", lty = c(1,2
```



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

```
Call:
survdiff(formula = Surv(time, cens) ~ group, data = leuk)

               N Observed Expected (O-E)^2/E (O-E)^2/V
group=6 MP    21        9     19.3      5.46      16.8
group=Placebo 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

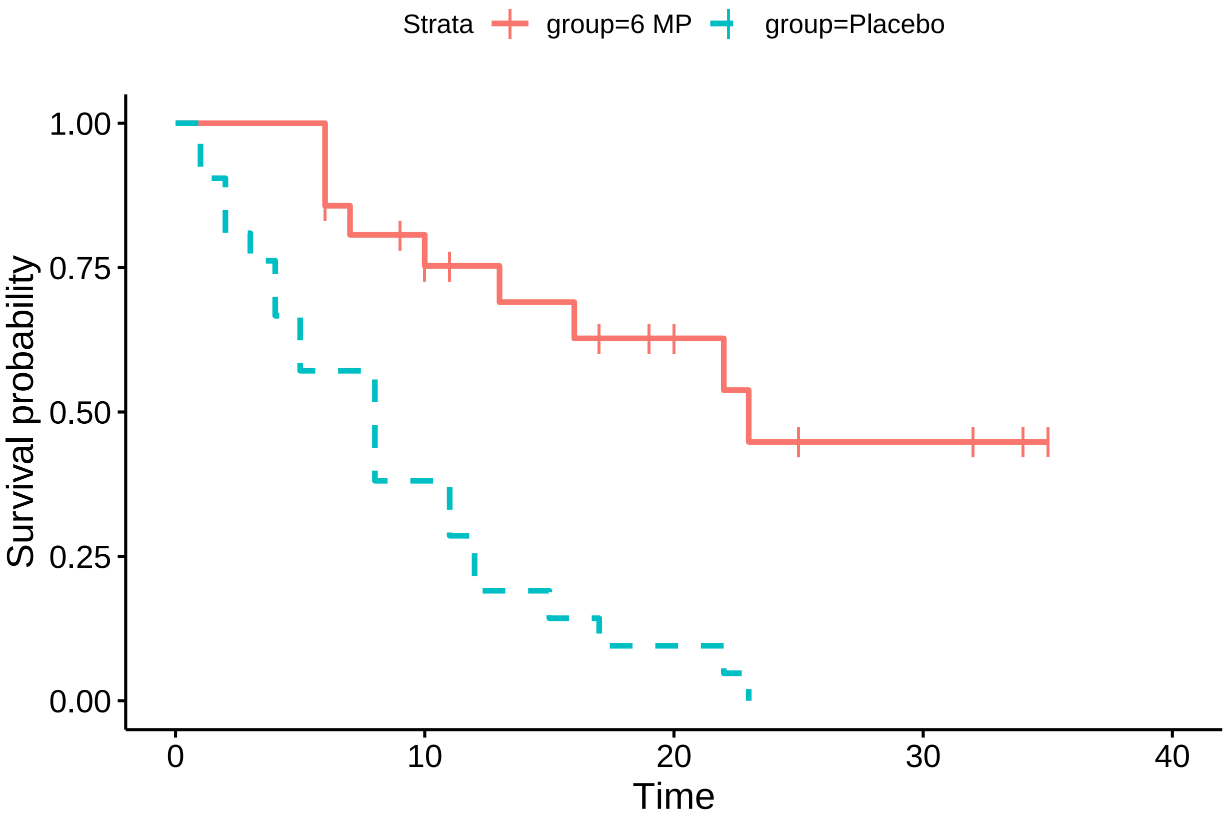
Definition: The survival function at time t, denoted S(t) is the probability of being eventfree at t; equivalently, the probability that the survival time is greater than t.

Interpretarion of Kaplan-Meier Curves

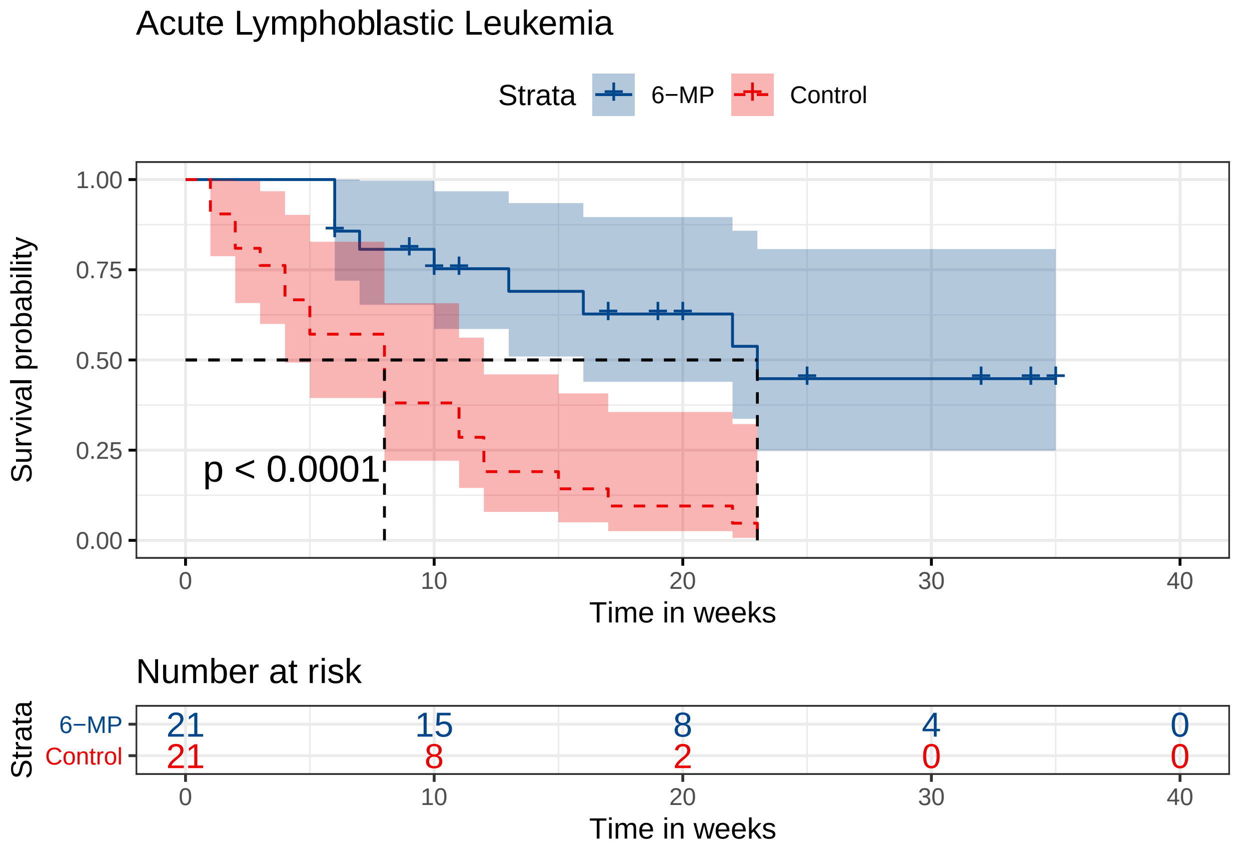
Plots of the Kaplan–Meier estimates of S(t), we can infer periods of high risk, when the survival curve descends rapidly, as well as periods of lower risk, when it remains relatively flat.

Fitting a survival model.

```
fit <- survfit(Surv(time, cens) ~ group, data = leuk)
ggsurvplot(fit, data = leuk, censor.shape="|", censor.size = 4, linetype = c(1,2))
```



```
# With ggplot using
ggsurvplot(
  fit,
  data = leuk,
  size = 0.5,                 # change line size
  linetype = c("solid", "dashed"), # different line type
  palette = c("lancet"), # color red, blue or custom palettes lancet
  title    = "Acute Lymphoblastic Leukemia", # plot main title
  xlab = "Time in weeks",    # customize X axis label.
  conf.int = TRUE,           # Add confidence interval
  pval = TRUE,               # Add p-value from log-rank test
  risk.table = TRUE,         # Add risk table
  risk.table.col = "strata",# Risk table color by groups
  legend.labs = c("6-MP", "Control"),     # Change legend labels
  risk.table.height = 0.25, # Useful to change when you have multiple groups
  surv.median.line = "hv",  # add the median survival pointer.
  ggtheme = theme_bw()      # Change ggplot2 theme
)
```



How is tha data organized for hese kind of problems?

```
leuk$group
```

```
 [1] "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"
 [8] "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"
[15] "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"    "6 MP"
[22] "Placebo" "Placebo" "Placebo" "Placebo" "Placebo" "Placebo" "Placebo"
[29] "Placebo" "Placebo" "Placebo" "Placebo" "Placebo" "Placebo" "Placebo"
[36] "Placebo" "Placebo" "Placebo" "Placebo" "Placebo" "Placebo" "Placebo"
```

```
leuk$time
```

```
 [1]  6  6  6  7 10 13 16 22 23  6  9 10 11 17 19 20 25 32 32 34 35  1  1  2  2
[26]  3  4  4  5  5  8  8  8  8 11 11 12 12 15 17 22 23
```

```
leuk$cens
```

```
 [1] 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[39] 1 1 1 1
```

Recidivism (the tendency of a convicted criminal to reoffend)

The Rossi data set in the CarData. Is a study of recidivism of 432 male prisoners, who were observed for a year after being released from prison. The following variables are included in the data:
`week`: week of the first arrest after release, or censoring time.
`arrest`: the event indicator, equal to 1 for those arrested during the period of the study and 0 for those who were not arrested.
`fin`: a factor, levels yes or no in the individual received finantial aid after release from prison.
`age`: in years at the time of release.

race: a factor with leves "black" and "other".
wexp: a factor "yes" if the individual had full-time work experience prior to incarceration or "no".
mar: a factor with leves "married" if te individual was married at the time of release and "not married" if he was not.
paro: a factor coded "yes" if the individual was released on parol and "no" if he was not.
prio: number of prior convictions.
educ: education, categorical variable coded numerically, with codes 2 (grade 6 or less), 3 (grades 6 through 9), 4 (grades 10 and 11), 5 (grade 12), or 6(some post-secondary).
*emp1-emp52: factors coded "yes" if the individual was employed in the corresponding week of the study.

```
# Cox proportional-hazards regression
# data("Rossi", package = "carData")
args(coxph)
```

```
function (formula, data, weights, subset, na.action, init, control,
    ties = c("efron", "breslow", "exact"), singular.ok = TRUE,
    robust, model = FALSE, x = FALSE, y = TRUE, tt, method = ties,
    id, cluster, istate, statedata, ...)
NULL
```

```
#
Rossi[1:5, 1:10]
```

```
  week arrest fin age  race wexp         mar paro prio educ
1   20      1  no  27 black   no not married  yes    3    3
2   17      1  no  18 black   no not married  yes    8    4
3   25      1  no  19 other  yes not married  yes   13    3
4   52      0 yes  23 black  yes     married  yes    1    5
5   52      0  no  19 other  yes not married  yes    3    3
```

```
# Cox model and estimation of model tests
mod.allison <- coxph(Surv(week, arrest) ~ fin + age + race + wexp + mar + paro + prio, data = Rossi)
summary(mod.allison)
```

```
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +
    mar + paro + prio, data = Rossi)

  n= 432, number of events= 114

                    coef exp(coef) se(coef)      z Pr(>|z|)
finyes          -0.37942   0.68426  0.19138 -1.983  0.04742 *
age             -0.05744   0.94418  0.02200 -2.611  0.00903 **
raceother       -0.31390   0.73059  0.30799 -1.019  0.30812
wexpyes         -0.14980   0.86088  0.21222 -0.706  0.48029
marnot married   0.43370   1.54296  0.38187  1.136  0.25606
paroyes         -0.08487   0.91863  0.19576 -0.434  0.66461
prio             0.09150   1.09581  0.02865  3.194  0.00140 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

               exp(coef) exp(-coef) lower .95 upper .95
finyes            0.6843     1.4614    0.4702    0.9957
age               0.9442     1.0591    0.9043    0.9858
raceother         0.7306     1.3688    0.3995    1.3361
wexpyes           0.8609     1.1616    0.5679    1.3049
```

```
marnot married    1.5430     0.6481    0.7300    3.2614
paroyes           0.9186     1.0886    0.6259    1.3482
prio              1.0958     0.9126    1.0360    1.1591

Concordance= 0.64  (se = 0.027 )
Likelihood ratio test= 33.27  on 7 df,   p=2e-05
Wald test            = 32.11  on 7 df,   p=4e-05
Score (logrank) test = 33.53  on 7 df,   p=2e-05
```

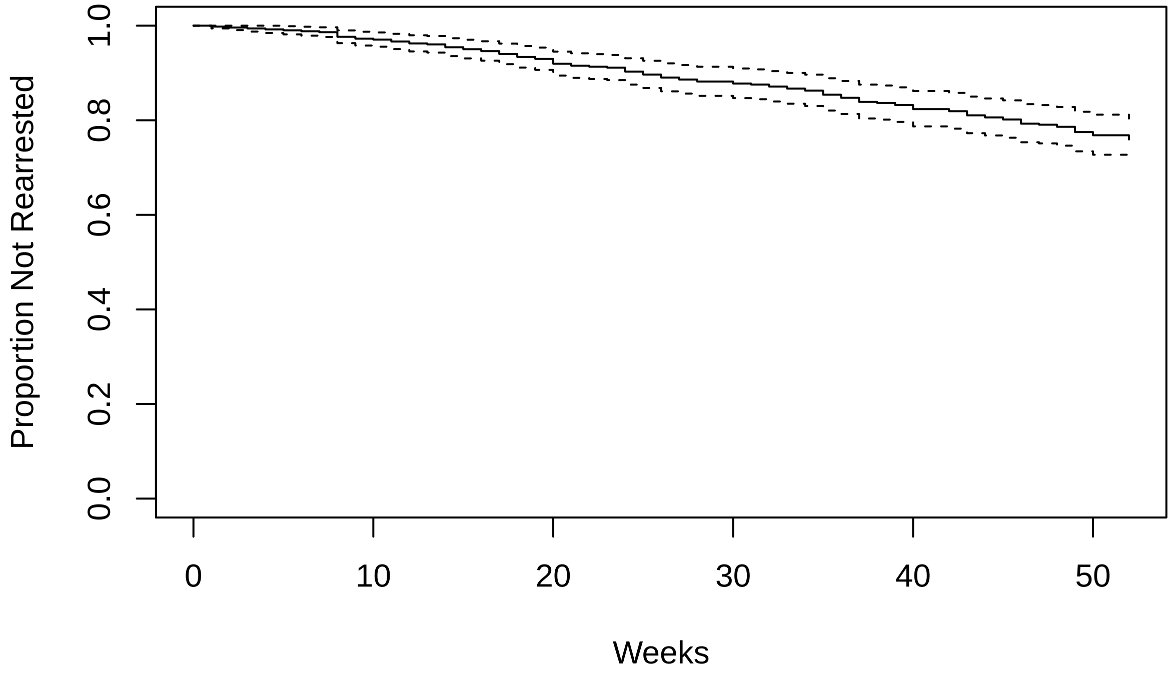
The Anova function (car)

The Anova function of car package has a method for "coxph" objects, by default estimates Type-II likelihood-ratio test for the terms of the Cox model

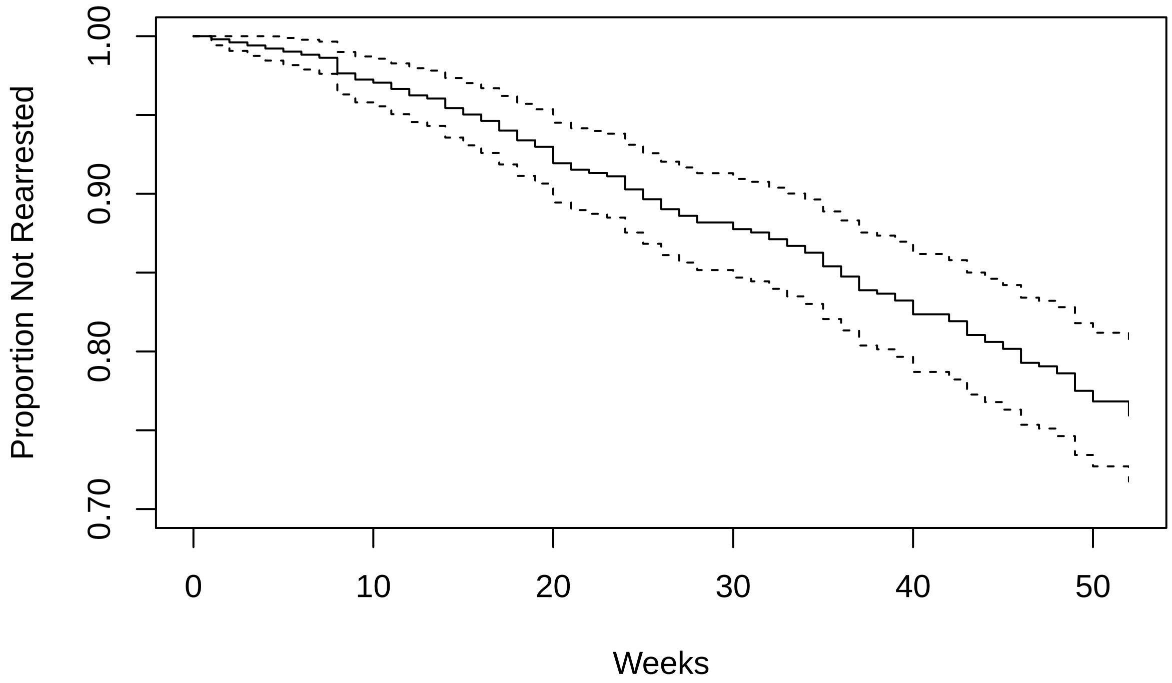
```
Anova(mod.allison)
```

```
Analysis of Deviance Table (Type II tests)
     LR Chisq Df Pr(>Chisq)
fin    3.9862  1   0.045874 *
age    7.9880  1   0.004709 **
race   1.1252  1   0.288812
wexp   0.5003  1   0.479352
mar    1.4312  1   0.231572
paro   0.1870  1   0.665450
prio   8.9766  1   0.002735 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

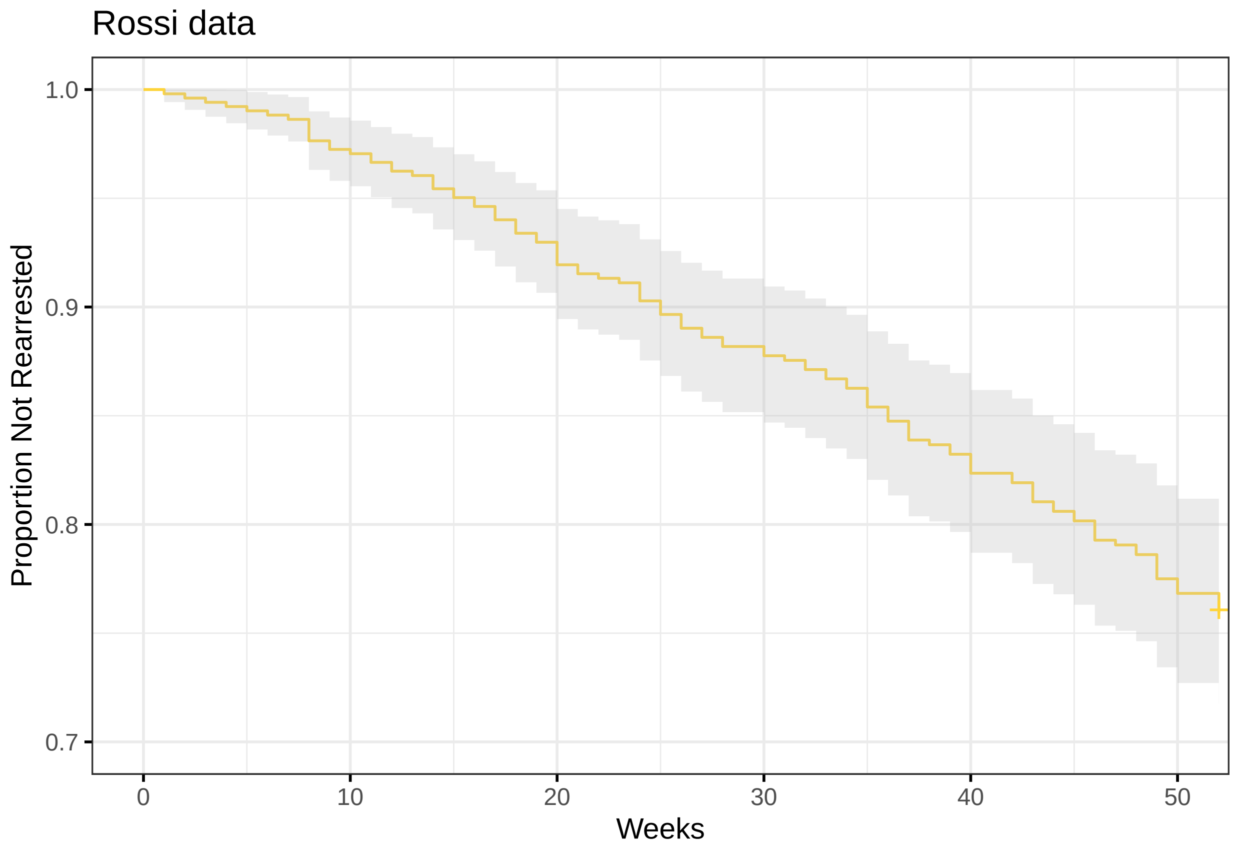
```
# Plots of the Cox model
plot(survfit(mod.allison), xlab="Weeks", ylab="Proportion Not Rearrested")
```



```
# now with better scale
plot(survfit(mod.allison), ylim = c(0.7, 1), xlab="Weeks", ylab="Proportion Not Rearrested")
```



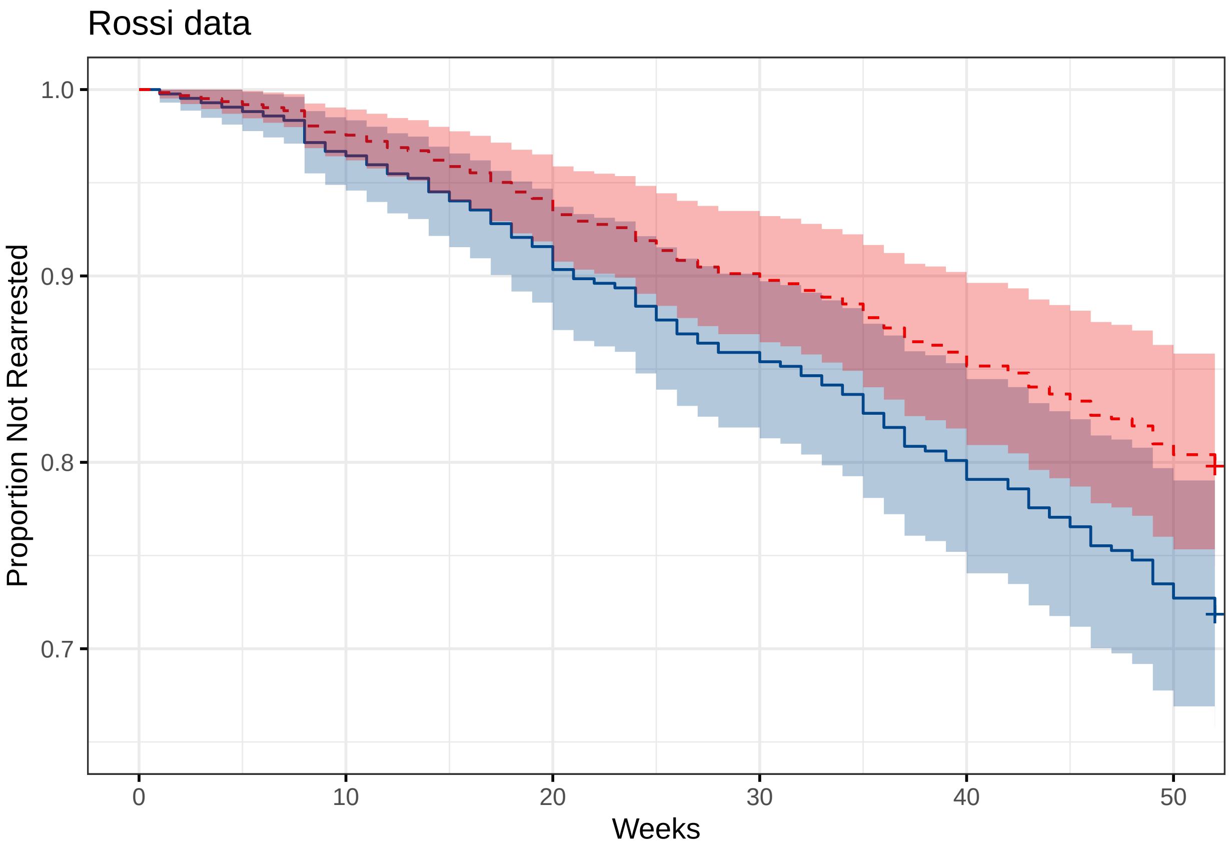
```
# With ggplot using ggsurvplot
#
ggsurvplot(
survfit(mod.allison),
data = Rossi,
size = 0.5,                      # change line size
linetype = c("solid","dashed"), # different line type
palette = "simpsons",              # color palette
title   = "Rossi data",          # plot main title
xlab = "Weeks",                  # customize X axis label.
ylab = "Proportion Not Rearrested", # customize Y axis label
ylim = c(0.7, 1),             # customize Y limits
conf.int = TRUE,             # Add confidence interval
pval = FALSE,                 # Add p-value from log-rank test
risk.table = FALSE,           # Add risk table
risk.table.col = "strata",   # Risk table color by groups
surv.median.line = "none",
legend = "none",
risk.table.height = 0.25,    # Useful to change when you have multiple groups
ggtheme = theme_bw()         # Change ggplot2 theme
)
```



```
# To study the fin help variable
Rossi.fin <- with (Rossi, data.frame(fin=c(0, 1), age=rep(mean(age), 2), race=rep(mean(race == "other")
                                     wexp=rep(mean(wexp == "yes"), 2), mar=rep(mean(mar == "not married
                                     paro=rep(mean(paro == "yes"), 2), prio=rep(mean(prio), 2)))

# plot(survfit(mod.allison, newdata=Rossi.fin), conf.int=TRUE, lty=c(1,2), ylim=c(0.6, 1), xlab="Weeks"
# legend("bottomleft", legend=c("fin = no", "fin = yes"), lty=c(1, 2), insert=0.02)

ggsurvplot(
  survfit(mod.allison, newdata=Rossi.fin),
  data = Rossi,
  size = 0.5,                     # change line size
  linetype = c("solid","dashed"), # different line type
  palette = "lancett",             # color palette
  title   = "Rossi data",         # plot main title
  xlab = "Weeks",                 # customize X axis label.
  ylab = "Proportion Not Rearrested", # customize Y axis label
  ylim = c(0.65, 1),           # customize Y limits
  conf.int = TRUE,            # Add confidence interval
  pval = FALSE,               # Add p-value from log-rank test
  risk.table = FALSE,         # Add risk table
  risk.table.col = "strata",  # Risk table color by groups
  surv.median.line = "none",
  legend = "none",
  risk.table.height = 0.25,   # Useful to change when you have multiple groups
  ggtheme = theme_bw()        # Change ggplot2 theme
  )
```


Rossi data
1.0
0.9
0.8
0.7
Proportion Not Rearrested
0
10
20
30
40
50
Weeks