HW 1 - Applied Survival Analysis, Due Thursday September 14th 11:59pm

Benjamin Panny

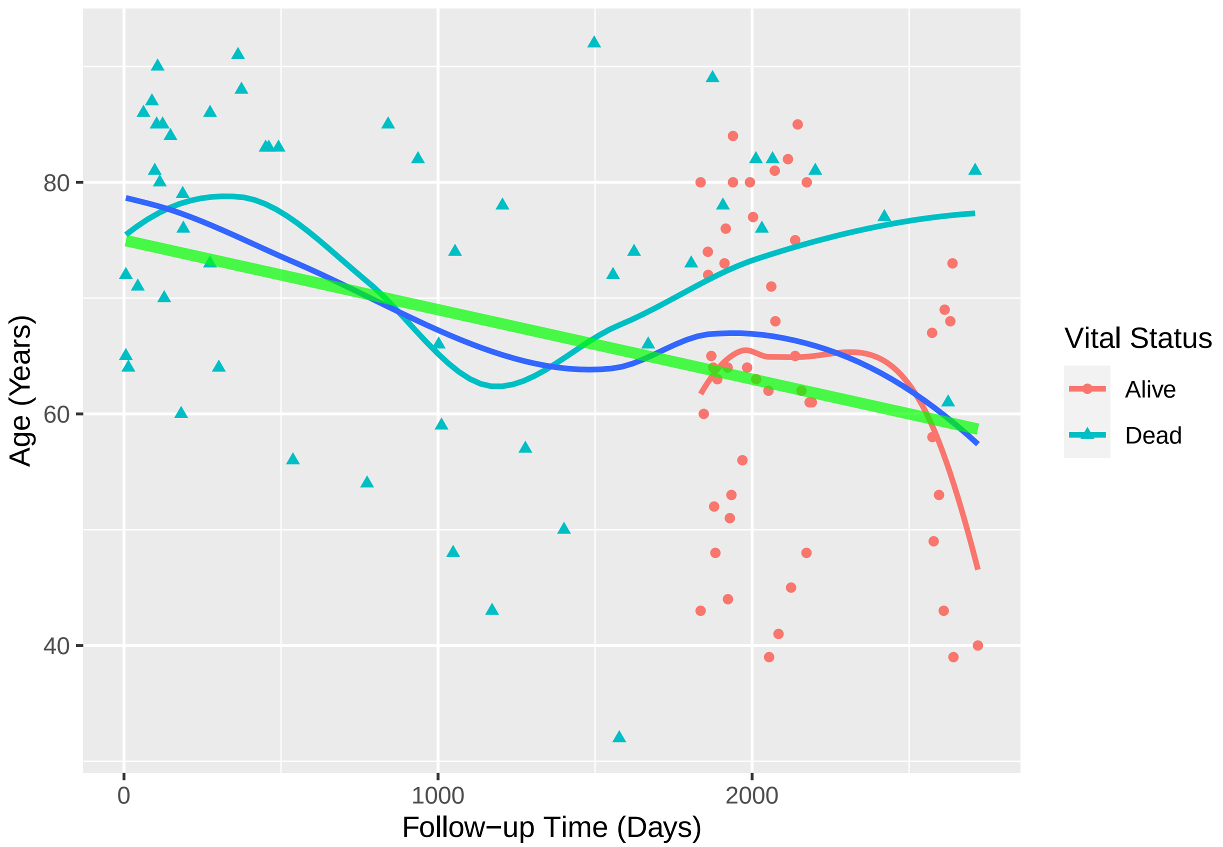
2023-09-15

1. (20 pts) Exercise 1 (a-d), textbook pages 13 and 15. The dataset "whas100.csv" can be found in the Homework/HW1 folder.

Using the data from the Worcester Heart Attack Study in Table 1.1, obtain a scatter plot of follow up time versus age. If possible, use the value of the vital status variable as the plotting symbol.

```r
whas %>%
  ggplot(aes(x = lenfol, y = age, shape = fstat, color = fstat)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  geom_smooth(aes(x = lenfol, y = age),
              inherit.aes = FALSE,
              se = FALSE) +
  geom_function(fun = ~ 75 - .006*(.x),
                inherit.aes = FALSE,
                color = "green",
                linewidth = 2,
                alpha = .7) +
  labs(shape = "Vital Status",
       color = "Vital Status",
       x = "Follow-up Time (Days)",
       y = "Age (Years)")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



(a) In what ways is the visual appearance of this plot different from a scatter plot in a typical linear regression setting?

All of the “Alive” Statuses have long follow-up times and fill up a square on the plot across virtually all of the variance in age. Most of the “Dead” Statuses have earlier follow-up times. Variance in age also fans over time, so a linear model may not have normal residuals; this pattern indicates that age may be protective.

(b) By eye, draw on the scatter plot from problem 1 (a) what you feel is the best regression function for a survival time regression model.

Plotted in Green by `geom_function()` and in blue as the LOESS curve.

(c) Is the regression function drawn in 1(b) a straight line? If not, then what function of age would you use to describe it?

I drew a straight line. A LOESS smoother applies a somewhat straight line. The lines are far more nonlinear when grouped by Vital Status at follow-up, both have roughly quadratic curvature.

(d) Is it possible to fit this model in your favorite software package with censored data?

Yes.

2. (10 pts) Exercise 3, textbook page 15.

The investigator of a large clinical trial would like to assess factors that might be associated with drop-out over the course of the trial. Describe what would be the event and which observations would be considered censored for such a study.

The event in this survival context is dropping out of the study, such as being unresponsive to study team contacts. A censored observation in this context could be if a participant completes the study (since we don't know when they would drop out should it continue forever), or a participant could drop out of the study for an unrelated reason, such as death or moving.

3. (20 pts) In January 1980 a large random sample of women age 18 and older in Allegheny County were contacted. Women without breast cancer were enrolled in an observational study and assessed every 3 years for the next 30 years, to determine the age at which they developed breast cancer. Describe the types of censoring and/or truncation that are represented by the following four individuals:

i. Enrolled at age 30, still without breast cancer at the year 30 assessment

Right Censoring because we do not know when they will develop cancer after the 30-year assessment, if at all.

ii. Enrolled at age 40, diagnosed with breast cancer at the year 15 assessment

No censoring or Truncation because their event was observed.

iii. Enrolled at age 50, still without breast cancer at the year 12 assessment, died in a boating accident at age 64

Left Censored because we do not know if and when they might have developed breast cancer in the next 18 years had they not died of an unrelated cause.

iv. Diagnosed with breast cancer at age 60 in 1976

Left Truncation because the study was not conducted until 1980.

4. (20 pts) The time to tumor development (in days) for rats exposed to a carcinogen follows a Weibull distribution with alpha = 2 and lambda = 0.001.

```r
weibull_survival <- function(x, lambda, alpha){
  exp(-lambda*(x^(alpha)))
}
weibull_probability <- function(x, lambda, alpha){
  lambda*alpha*(x^(alpha - 1))*exp(-lambda*(x^(alpha)))
}
weibull_hazard <- function(x, lambda, alpha){
    probability <-  weibull_probability(x, lambda, alpha)
    survival <- weibull_survival(x, lambda, alpha)
    hazard <- probability / survival
    return(hazard)
}
```

a. (10 pts) What is the probability a rat will be tumor-free at 30 days? 45 days? 60 days?

30 days: 0.4065697 (`weibull_survival(30, lambda = 0.001, alpha = 2)`)

45 days: 0.1319938 (`weibull_survival(45, lambda = 0.001, alpha = 2)`)

60 days: 0.0273237 (`weibull_survival(60, lambda = 0.001, alpha = 2)`)

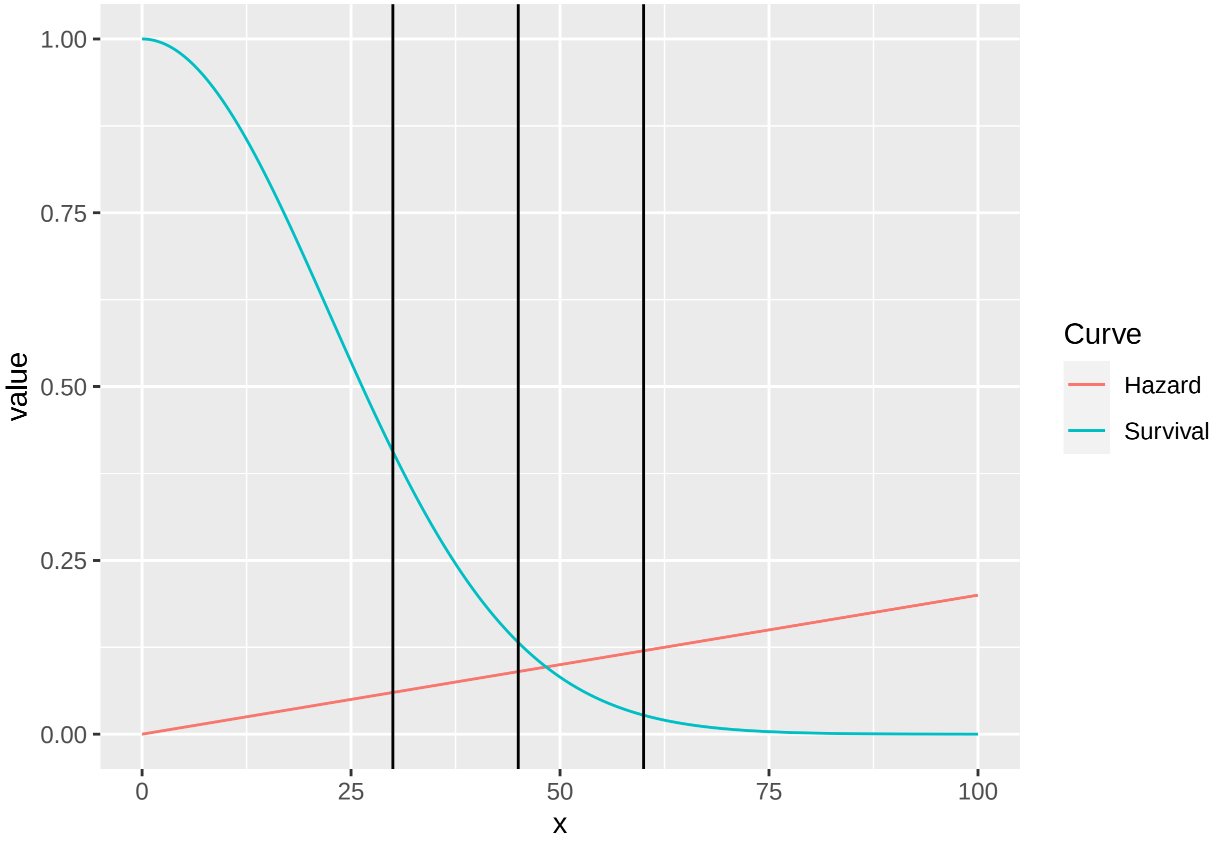
b. (10 pts) Find the hazard rate of the time to tumor appearance at 30 days, 45 days, and 60 days.

30 days: 0.06 (`weibull_hazard(30, lambda = 0.001, alpha = 2)`)

45 days: 0.09 (`weibull_hazard(45, lambda = 0.001, alpha = 2)`)

60 days: 0.12 (`weibull_hazard(60, lambda = 0.001, alpha = 2)`)

```r
l <- 0.001; a <- 2
x <- seq(0, 100, by = .01)
tibble(x = x, Survival = weibull_survival(x, l, a), Hazard = weibull_hazard(x, l, a)) %>%
  pivot_longer(!x,
               names_to = "Curve") %>%
  ggplot(aes(x = x, y = value, color = Curve)) +
  geom_line() +
  geom_vline(xintercept = c(30, 45, 60))
```



5. (30 pts) Listed below are values of survival time in years for 6 males and 6 females from the WHAS100 study. Right-censored times are denoted by a “+” as a superscript.

Table 1: Females

Interval	Conditional Probability (K-M Estimates of S(t))
0 <= t < 0.4	1.0000000
0.4 <= t < 1.2	0.8333333
1.2 <= t < 4.3	0.6666667
4.3 <= t < 4.9	0.5000000
4.9 <= t < 5.0	0.3333333
5.0 <= t < 5.1	0.1666667
5.1 <= t	0.1666667

Males: 1.2, 3.4, 5.0+, 5.1, 6.1, 7.1 Females: 0.4, 1.2, 4.3, 4.9, 5.0, 5.1+

Using these data: a. (20 pts) compute the following by hand the Kaplan-Meier estimate of the survival function for each sex group.

```r
whas6 <- tibble(lenfol = c(1.2, 3.4, 5.0, 5.1, 6.1, 7.1,
                 0.4, 1.2, 4.3, 4.9, 5.0, 5.1),
       fstat = c(1,1,0,1,1,1,1,1,1,1,1,0),
       sex = c(rep(c("male", "female"), each = 6))) %>%
  arrange(sex, lenfol)

tibble(Interval = c("0 <= t < 0.4",
                    "0.4 <= t < 1.2",
                    "1.2 <= t < 4.3",
                    "4.3 <= t < 4.9",
                    "4.9 <= t < 5.0",
                    "5.0 <= t < 5.1",
                    "5.1 <= t"),
       `Conditional Probability (K-M Estimates of S(t))` = c(1,
                                                             1*5/6,
                                                             1*5/6*4/5,
                                                             1*5/6*4/5*3/4,
                                                             1*5/6*4/5*3/4*2/3,
                                                             1*5/6*4/5*3/4*2/3*1/2,
                                                             1*5/6*4/5*3/4*2/3*1/2*1/1)) %>%
  kable(caption = "Females") %>%
  kable_styling()
```

```r
tibble(Interval = c("0 <= t < 1.2",
                    "1.2 <= t < 3.4",
                    "3.4 <= t < 5",
                    "5 <= t < 5.1",
                    "5.1 <= t < 6.1",
                    "6.1 <= t < 7.1",
                    "7.1 <= t"),
       `Conditional Probability (K-M Estimates of S(t))` = c(1,
                                                             1*5/6,
                                                             1*5/6*4/5,
                                                             1*5/6*4/5*4/4,
                                                             1*5/6*4/5*4/4*3/4,
                                                             1*5/6*4/5*4/4*3/4*2/3,
```

Table 2: Males

Interval	Conditional Probability (K-M Estimates of S(t))
0 <= t < 1.2	1.0000000
1.2 <= t < 3.4	0.8333333
3.4 <= t < 5	0.6666667
5 <= t < 5.1	0.6666667
5.1 <= t < 6.1	0.5000000
6.1 <= t < 7.1	0.3333333
7.1 <= t	0.0000000

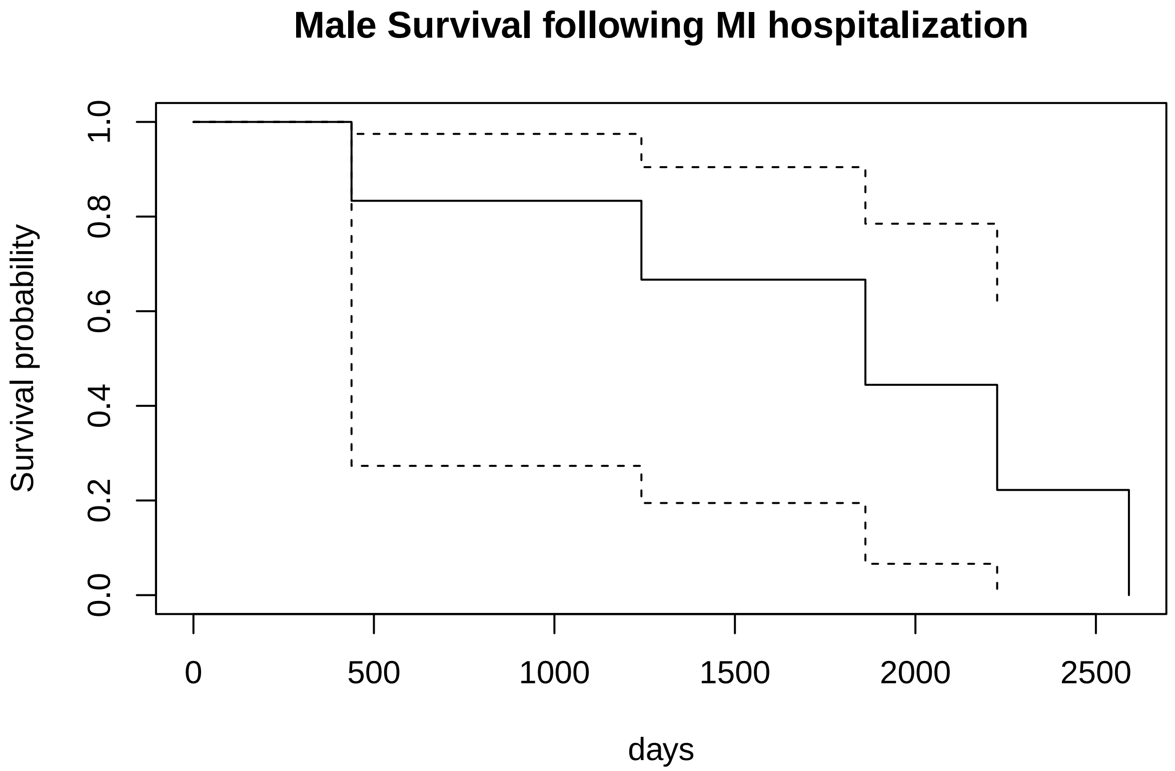
```
                                                                     1*5/6*4/5*4/4*2/3*1/2*0/1)) %>%
  kable(caption = "Males") %>%
  kable_styling()
```

b. (10 pts) Use SAS or R to get pointwise 95 percent confidence intervals for the survival functions estimated in problem (a). Use log-log transformation for the variance estimate.

```
# read in WHAS100 data
library(survival)
# produce K-M estimate with 95% log-log based CI
KM.whas6_male <- survfit(Surv(lenfol*365,fstat)~1,
                         data=whas6 %>% filter(sex == 'male'),
                         conf.int=0.95, error = "greenwood", conf.type="log-log")
summary(KM.whas6_male)
```

```
## Call: survfit(formula = Surv(lenfol * 365, fstat) ~ 1, data = whas6 %>%
##     filter(sex == "male"), error = "greenwood", conf.int = 0.95,
##     conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   438      6       1    0.833   0.152      0.27312        0.975
##  1241      5       1    0.667   0.192      0.19462        0.904
##  1861      3       1    0.444   0.222      0.06619        0.785
##  2226      2       1    0.222   0.192      0.00957        0.615
##  2592      1       1    0.000     NaN           NA           NA
```

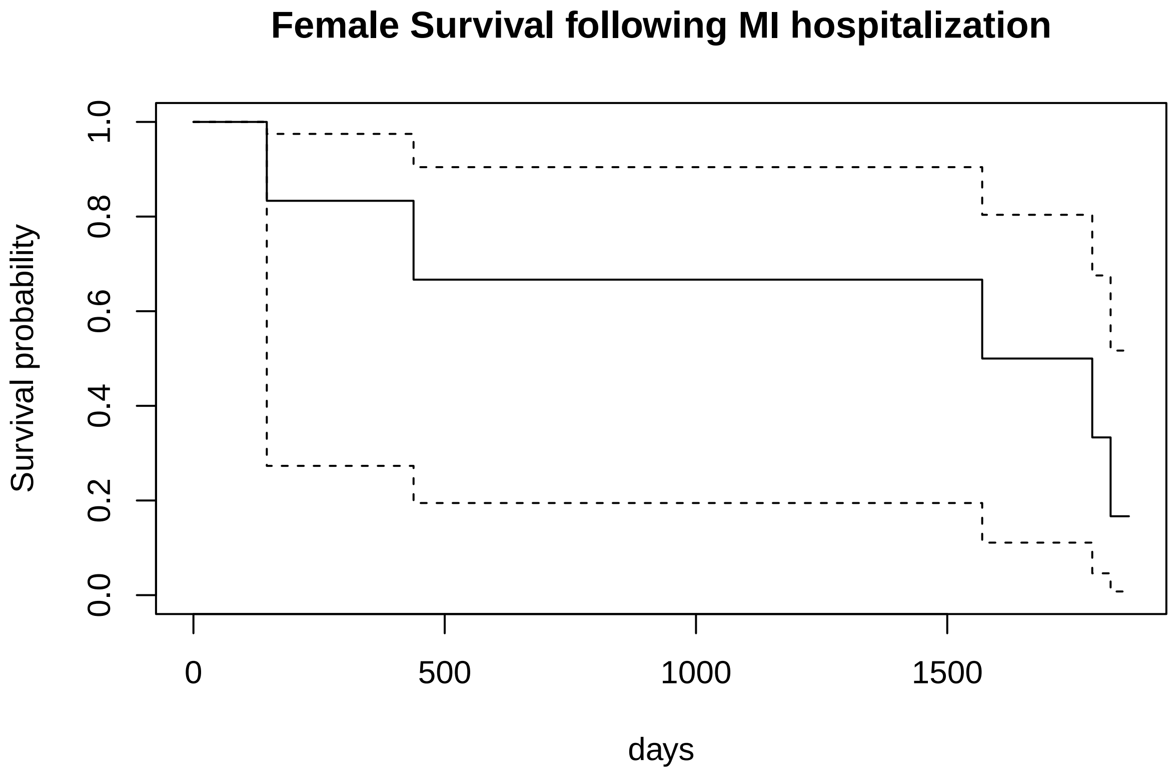
```
# plot K-M estimate and its associated CI
plot(KM.whas6_male,
     main="Male Survival following MI hospitalization",
     xlab="days",
     ylab="Survival probability")
```



```
KM.whas6_female <- survfit(Surv(lenfol*365,fstat)~1,
                           data=whas6 %>% filter(sex == 'female'),
                           conf.int=0.95, error = "greenwood", conf.type="log-log")
summary(KM.whas6_female)
```

```
## Call: survfit(formula = Surv(lenfol * 365, fstat) ~ 1, data = whas6 %>%
##     filter(sex == "female"), error = "greenwood", conf.int = 0.95,
##     conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   146      6       1    0.833   0.152      0.27312        0.975
##   438      5       1    0.667   0.192      0.19462        0.904
##  1570      4       1    0.500   0.204      0.11095        0.804
##  1789      3       1    0.333   0.192      0.04608        0.676
##  1825      2       1    0.167   0.152      0.00772        0.517
```

```
# plot K-M estimate and its associated CI
plot(KM.whas6_female,
     main="Female Survival following MI hospitalization",
     xlab="days",
     ylab="Survival probability")
```


Female Survival following MI hospitalization
Survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
days