Problem Set 4

Applied Stats II

Due: April 12, 2024

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub in `.pdf` form.
- This problem set is due before 23:59 on Friday April 12, 2024. No late assignments will be accepted.

Question 1

We're interested in modeling the historical causes of child mortality. We have data from 26855 children born in Skellefteå, Sweden from 1850 to 1884. Using the "child" dataset in the `eha` library, fit a Cox Proportional Hazard model using mother's age and infant's gender as covariates. Present and interpret the output.

```
######################
# Problem 1
######################
```

```r

# load data
dat <- child # use the child data set in eha

# from the eha network codebook,
# https://cran.r-project.org/web/packages/eha/eha.pdf - page 13
# we can see the instruction
# - sex: Sex
# - m.age: mother age

child_surv <- with(child, Surv(enter, exit, event))
km <- survfit(child_surv ~ 1, data = child)
summary(km, times = seq(0, 15, 1))

km_data <- data.frame(time = km$time, surv = km$surv)

# get CI
lower_ci <- km$lower
upper_ci <- km$upper

# visualisation the curve
pdf("Kaplan-Meier.pdf")
ggplot(km_data, aes(x = time, y = surv)) +
  geom_step() +
  geom_ribbon(aes(ymin = lower_ci, ymax = upper_ci),
              fill="#3F72AF", color="black", linetype =
                 "dashed", alpha = 0.5) +
  labs(title = "Kaplan-Meier Plot with Confidence Interval", x
     = "Years", y = "Survival Probability") +
  ylim(0.7, 1) +
  theme_bw()
dev.off()

COX <- coxph(child_surv ~ sex + m.age, data = dat)
summary(COX)
```

Firstly we will get a Kaplan-Meier curve:

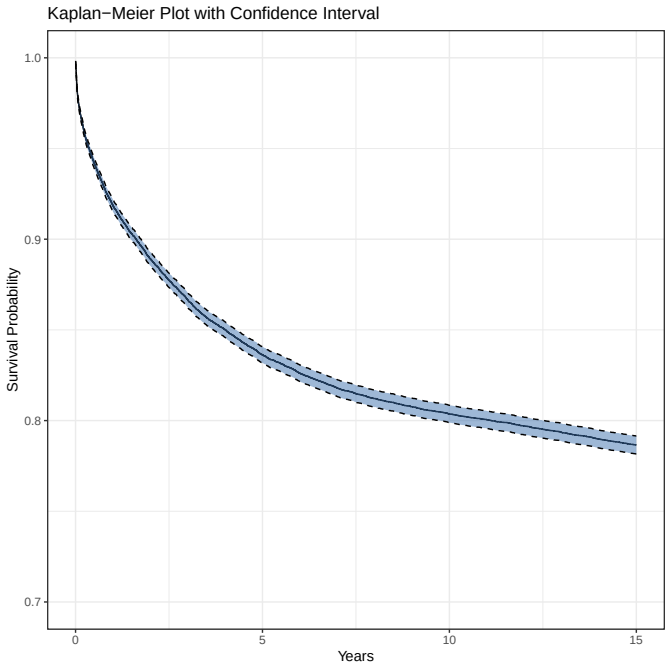

Figure 1: Kaplan-Meier Plot

When near the 0 year, all individual haven't experience the event, so the probability of living is 1 and where the upper limit and lower limit of CI are also 1.

With time goes by, the probability of living is decreasing and the CI is also getting wider, which means the uncertain of living probability estimation is increasing.

And we can get our model here:

```
Call:
coxph(formula = child_surv ~ sex + m.age, data = dat)

  n= 26574, number of events= 5616

               coef exp(coef)  se(coef)      z Pr(>|z|)
sexfemale -0.082215  0.921074  0.026743 -3.074 0.002110 **
m.age      0.007617  1.007646  0.002128  3.580 0.000344 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
sexfemale    0.9211     1.0857     0.874    0.9706
m.age        1.0076     0.9924     1.003    1.0119

Concordance= 0.519  (se = 0.004 )
Likelihood ratio test= 22.52  on 2 df,   p=1e-05
Wald test            = 22.52  on 2 df,   p=1e-05
Score (logrank) test = 22.53  on 2 df,   p=1e-05
```

Here are some key results:

- For the exp coefficient of the sexfemale, we can interpret that compared with male, female related to a 0.92 increasing in average of the probability in risk.
- For the exp coefficient of the m.age, we can interpret that 1 year increase in age is associated with an increase of 1.01 on average for the probability in risk.