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.

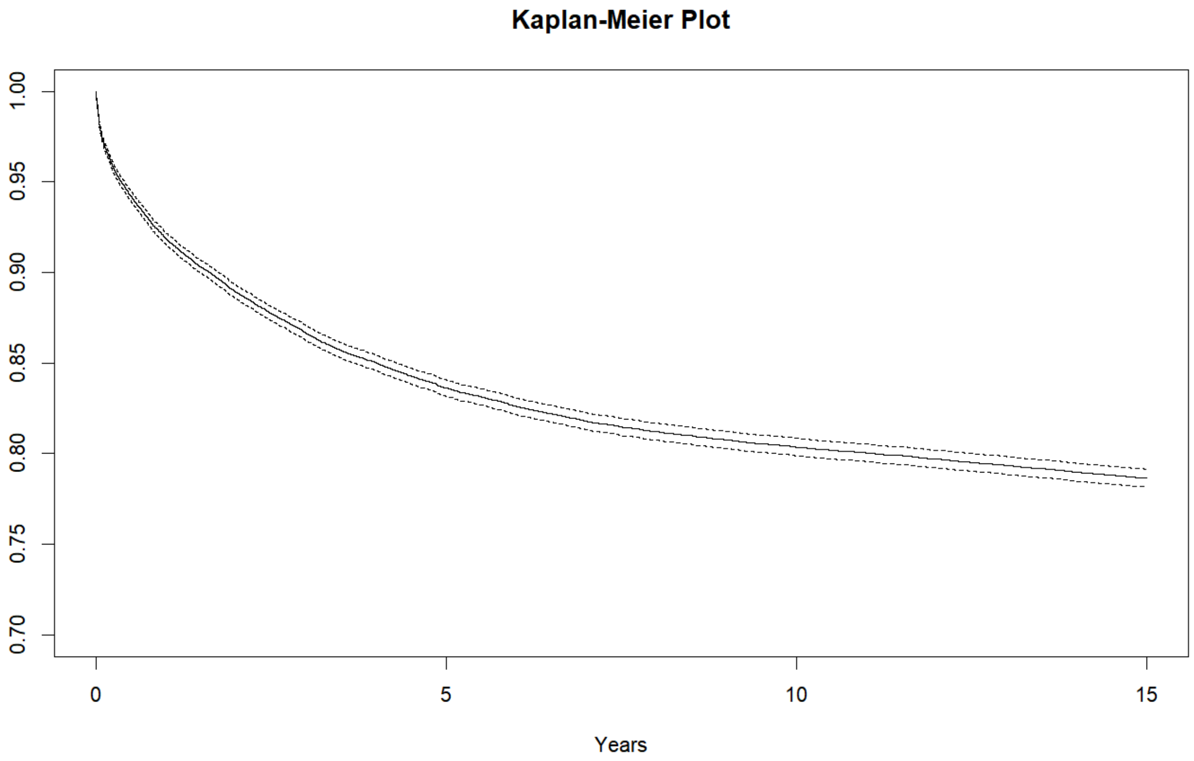
```r
# Download the dataset
library('eha')
library(survival)
install.packages(strag)
library(stargazer)
library(ggplot2)
data(child)
print(child)

# Create a object with time-to-live and event state information
child_surv <- with(child, Surv(enter, exit, event))

# The Kaplan-Meier method was used to estimate survival functions and plot survival curves.
km <- survfit(child_surv ~ 1, data = child)
summary(km, times = seq(0, 15, 1))
```

```
plot(km, main = "Kaplan-Meier Plot", xlab = "Years", ylim = c(0.7, 1))


# Fita Cox Proportional Hazard model using mother's age and infant's gender
cox <- coxph(child_surv ~ m.age + sex, data = child)
summary(cox)
stargazer(cox)
```



According to the picture, the survival rate decreases fastest from 0 to 5 years from 1 to about 0.8, which means that the survival risk is high during this period; the curve decreases slowly from 5 to 15 years, which means that the survival risk is low during this period.

Table 1:

	Dependent variable:
	child_surv
m.age	0.008^{***}
	(0.002)
sexfemale	-0.082^{***}
	(0.027)
Observations	26,574
R^2	0.001
Max. Possible R^2	0.986
Log Likelihood	$-56{,}503.480$
Wald Test	22.520^{***} (df = 2)
LR Test	22.518^{***} (df = 2)
Score (Logrank) Test	22.530^{***} (df = 2)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient of m.age is 0.007617, which means that for every unit increase in maternal age, the hazard ratio increases by 1.007646 times, and the p-value of the coefficient is less than and not equal to 0.01, so the coefficient is statistically significant.

The coefficient of sexfemale is -0.082215, which means that compared with the reference group, the risk ratio of women is reduced to 0.921074 times, and the p-value of the coefficient is less than and not equal to 0.01, so the coefficient is statistically significant.