Problem Set 4

Applied Stats II

Due: April 16, 2023

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 Sunday April 16, 2023. 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.

My Answer - Question 1

Looking at the variables in the imported data(infants)

- enter - Age (in days) of case when its mother died
- exit - Age (in days) at death or right censoring (at age 365 days)
- event - Follow up ends with death (1) or right censoring (0)
- mother - dead for cases, alive for controls
- age - Mother's age at infant's birth

- sex - The infant’s sex
- parish - Birth parish, either Nedertornea or not Nedertornea (other)
- civst - Civil status of mother, married or unmarried
- ses - Socio economic status of mothe, either farmer or not farmer (other)
- year - Year of birth of the infant

In order to fit a Cox Proportional Hazard Model the coxph() function from the ’survival’ package is required. This needs to be in a Surv() format which represents time to event data. The first argument is time variable.

Testing with Weibull proportional hazard model. Proportional hazards model with parametric baseline hazard(s). Allows for stratification with different scale and shape in each stratum, and left truncated and right censored data. Create a new variable called ”gender” in the ”infants” data frame, and assign it values based on the ”sex” variable. Specifically, the first command sets all the values of ”gender” to missing (NA) for all observations in the ”infants” data frame. The second command sets the value of ”gender” to 0 for all observations in the ”infants” data frame where the value of ”sex” is equal to ’boy’. This means that all male infants in the data frame will have a value of 0 in the ”gender” variable. The third command sets the value of ”gender” to 1 for all observations in the ”infants” data frame where the value of ”sex” is equal to ’girl’. This means that all female infants in the data frame will have a value of 1 in the ”gender” variable. In summary, these commands create a new variable ”gender” in the ”infants” data frame and assign values of 0 and 1 to it based on the ”sex” variable. This is a common technique used to recode categorical variables with two categories into a binary format for statistical analysis. Kaplan Meier survival curve for the ”infants s” variable in the ”infants” data frame and provide a summary of the survival probabilities at specific time points. This is a common technique used in survival analysis to estimate the probability of an event (e.g., death, failure) over time, based on a set of predictor variables.

```r


# Read in the data and have a look at it
data(infants)
glimpse(infants)
infants
table(infants)

# inspecting the data
str(infants)
names(infants)

```

```r
table(infants$sex)
# girl  boy
#  34   71

prop.table(table(infants$sex))
# girl       boy
# 0.3238095 0.6761905

# recode
infants$gender <- NA
infants$gender[infants$sex=='boy'] <- 0
infants$gender[infants$sex=='girl'] <- 1

# histogram to look at the visual of the age and sex variables
plot_ly(infants, x = ~age, color = ~sex) %>%
  add_histogram()


fit <- coxreg(Surv(enter, exit, event) ~ strata(stratum) + mother, data
              = infants)
fit
fit.w <- phreg(Surv(enter, exit, event) ~ mother + parish + ses, data =
                  infants)
summary(fit.w) ## Weibull proportional hazards model.
plot(fit.w)

coef(fit.w) # for extracting the coefficients from the model


infants_s <- with(infants, Surv(enter, exit, event))
```

The Kaplan-Meier method is a non parametric statistic that allows you to estimate the survival function.

```r


# The Kaplan-Meier method is a non parametric statistic that allows you to
    estimate the survival function

# plotting Kaplan-Meier method
# The resulting object "kaplan" is a "survfit" object that stores the Kaplan-
    Meier estimates of the survival probabilities,
# as well as the number of observations and the number of events at each time
    point.
# The object can be used for further analysis or plotting.


kaplan <- survfit(infants_s ~ 1, data = infants)
summary(kaplan, times = seq(0, 15, 1))

plot(kaplan, main = "Kaplan-Meier Method", xlab = "Days", ylim = c(0.5, 1))
autoplot(kaplan)
```

```r

# plotting the sex covariates using the infants gender , girl and boy
kaplan_sex <- survfit(infants_s ~ sex, data = infants)
autoplot(kaplan_sex)

# running the Cox Proportional Hazard Model, using mothers age and infants gender as the covariates
cox <- coxph(Surv(enter, exit, event) ~ age + sex, data = infants)
summary(cox)
```

Interpretation

```r


############ Interpretation

# there is a .485 decrease in the expected log of the hazard for male babies in comparison to female babies,
# while holding the age of the moether constant

# there is .04 decrease in the expected log of of the hazrd each time the mothers age increases by 1 year or 1 unit increase
# while holding the sex of infant constant

# coef    exp(coef) se(coef)                 z Pr(>|z|)
# age    -0.04044   0.96037  0.04507 -0.897    0.370
# sexboy -0.48518   0.61559  0.44224 -1.097    0.273


### hazard ratio for male babies is .61 compared to that of female babies or 62 male babies die for every 100 female babies
exp(coef(cox))
# age       sexboy
# 0.9603673 0.6155879


stargazer(cox)


######################################## Dependent variable  is : infant_s
# testing the out of the coefficients
coef(cox_s)
# age       sexboy
# -0.04043946 -0.48517752

exp(coef(cox_s))
# age     sexboy
# 0.9603673 0.6155879


```

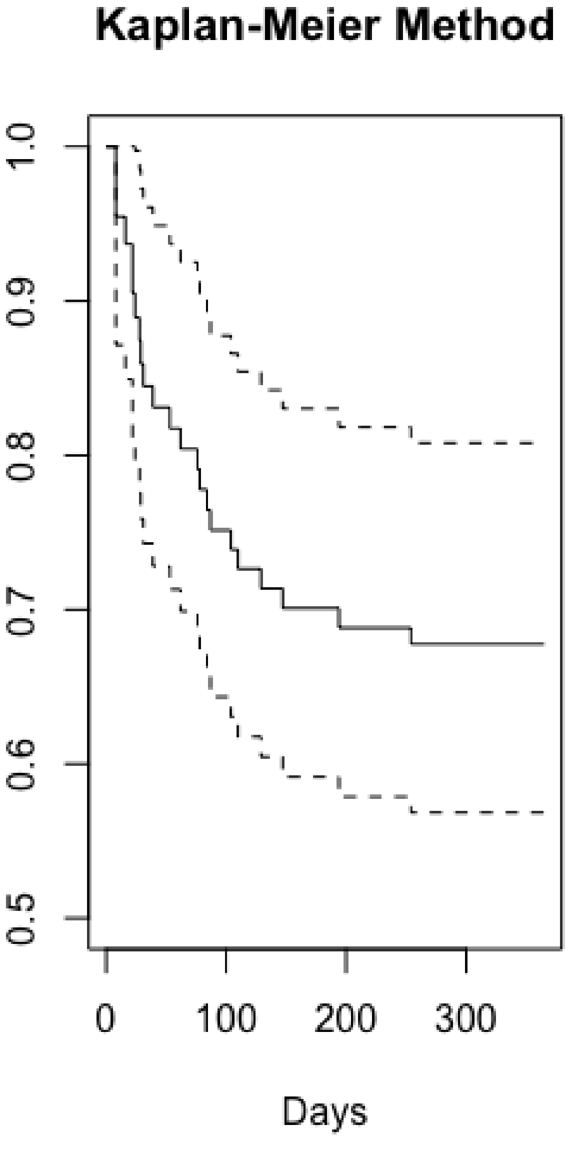

Figure 1: Kaplan-Meier Method

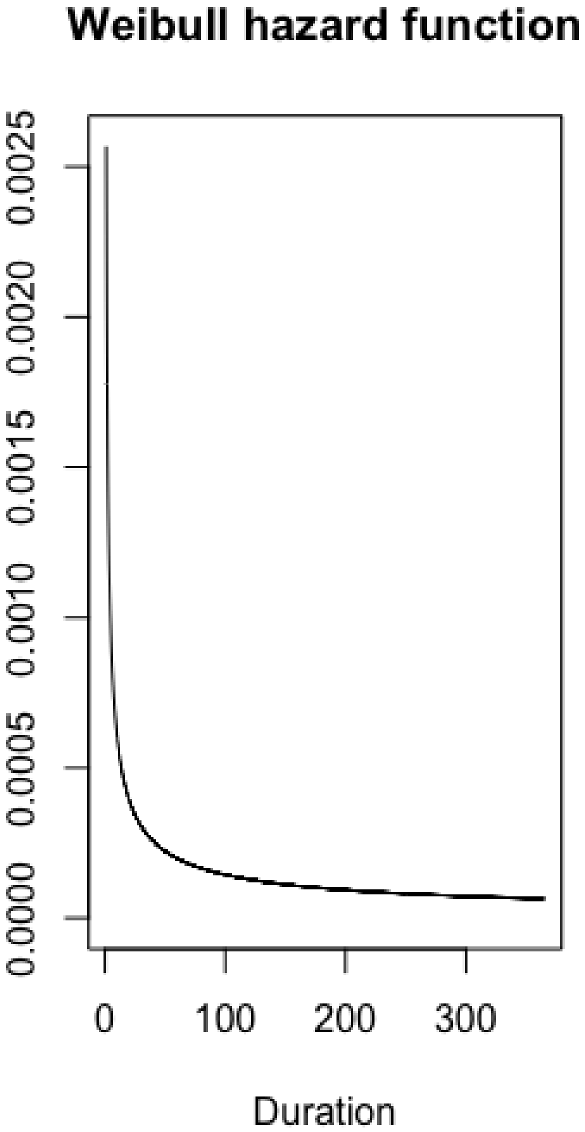

Figure 2: Weibull hazard function

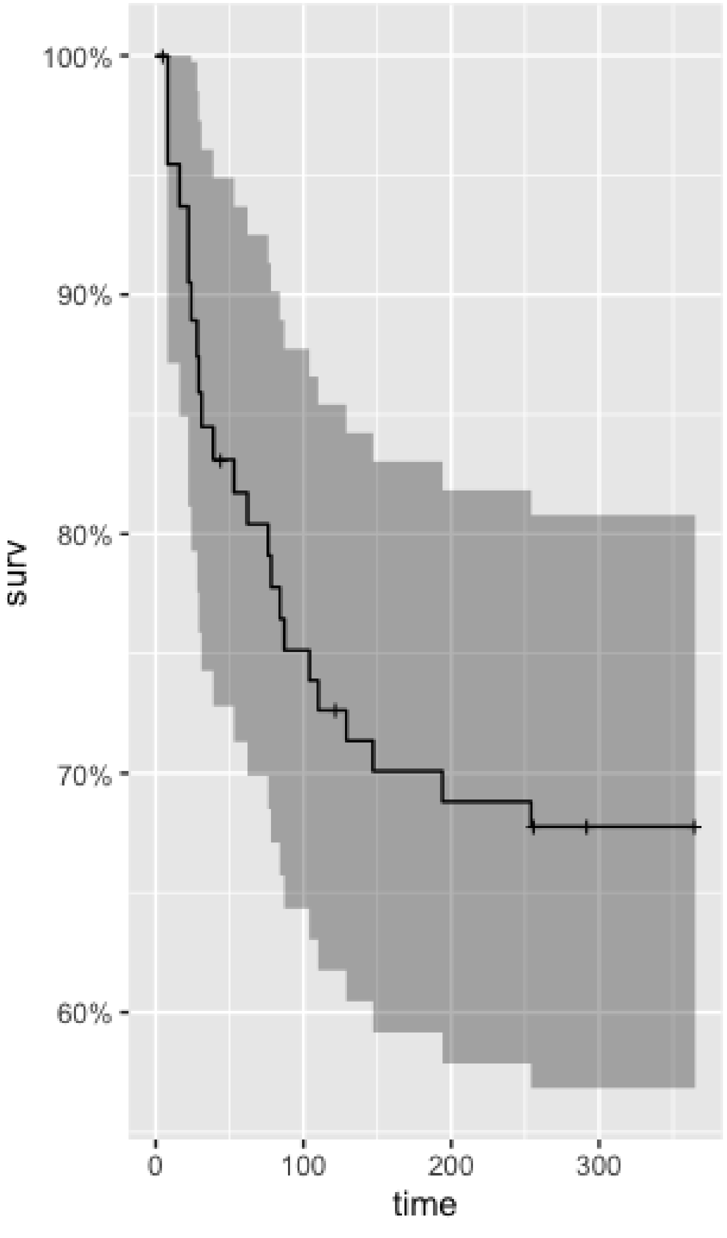

Figure 3: Kaplan-Meir Plot

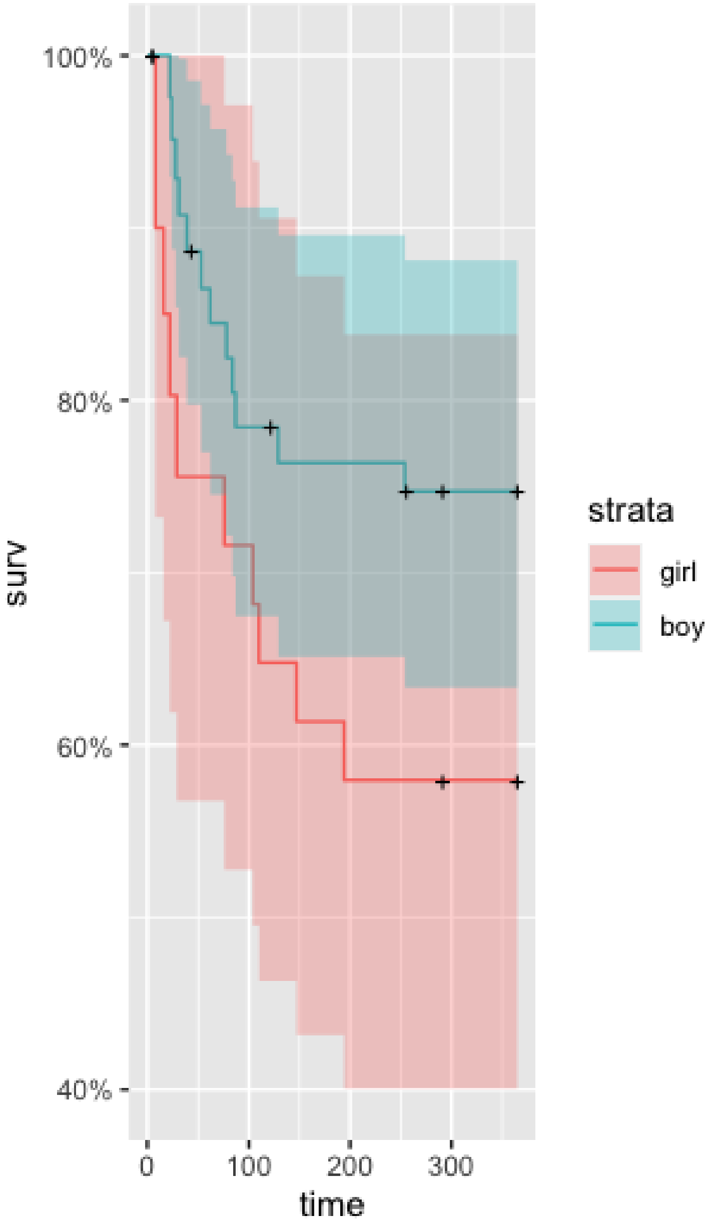

Figure 4: Kaplan-Meir Plot

```r
# this command (below) performs a Cox proportional hazards regression analysis to estimate the hazard
# rate of an event over time, while controlling for the effects of age and sex in the "infants" data frame.
# This is a common technique used in survival analysis to model the effects of predictor variables on
# the risk of an event, and to identify factors that may be associated with increased or decreased risk.

cox.w <- phreg(Surv(enter, exit, event) ~ age + sex, data = infants)
summary(cox.w)
plot(cox.w)

# Covariate                Mean        Coef       Rel.Risk    S.E.     LR p
# age                     27.127      -0.050      0.951      0.045    0.2376
# sex                                                                 0.4029
# girl                   0.317       0          1 (reference)
# boy                    0.683      -0.375      0.687      0.444

# Events                        21
# Total time at risk             21616
# Max. log. likelihood         -154.85
# L3 R test statistic             2.10
```