Problem Set 4 _ Stats II

Dongli Wu_21362988

4/3/2022

Instruction

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 class on Monday April 4, 2022. No late assignments will be accepted. - Total available points for this homework is 80.

Question 1

We're interested in modeling the historical causes of infant mortality. We have data from 5641 first-born in seven Swedish parishes 1820-1895. Using the "infants" 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
# import data
data(infants)
# have a look at some details about this infants data information
help(infants)
```

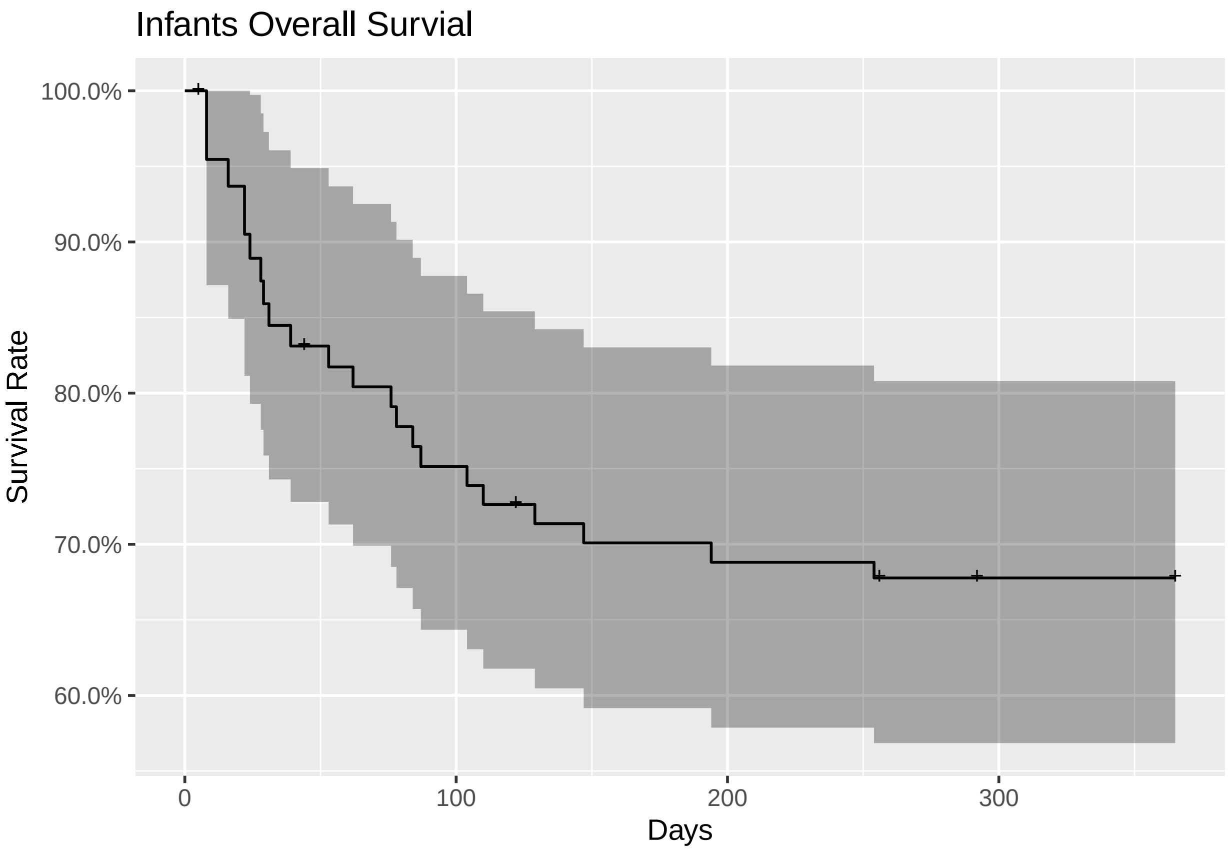
A data frame with 80 rows and five variables.

- stratum: Triplet No. Each triplet consist of one infant whose mother died (a case), and two controls, i.e, infants whose mother did not die. Matched on covariates below.
- 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.
- civst: Civil status of mother, married or unmarried.
- ses: Socio-economic status of mother, either farmer or not farmer.
- year: Year of birth of the infant.

Run a Kaplan-Meier plot for overall survival

```r
# build a survival object
infants_surv <- with(infants, Surv(enter, exit, event))
```

```
# Kaplan-Meier plot
km <- survfit(infants_surv ~ 1, data = infants)
autoplot(km, main = "Infants Overall Survial", xlab = "Days", ylab = "Survival Rate")
```



```
# use summary function to get a general discription
summary(km)
```

```
## Call: survfit(formula = infants_surv ~ 1, data = infants)
##
##  time n.risk n.event censored survival std.err lower 95% CI upper 95% CI
##     8     22       1        2    0.955  0.0444        0.871        1.000
##    16     54       1        0    0.937  0.0470        0.849        1.000
##    22     59       2        0    0.905  0.0505        0.811        1.000
##    24     57       1        0    0.889  0.0520        0.793        0.997
##    28     59       1        0    0.874  0.0533        0.776        0.985
##    29     58       1        0    0.859  0.0544        0.759        0.973
##    31     60       1        0    0.845  0.0554        0.743        0.961
##    39     62       1        0    0.831  0.0561        0.728        0.949
##    53     60       1        1    0.817  0.0569        0.713        0.937
##    62     62       1        0    0.804  0.0575        0.699        0.925
##    76     61       1        0    0.791  0.0580        0.685        0.913
##    78     60       1        0    0.778  0.0585        0.671        0.901
##    84     59       1        0    0.765  0.0590        0.657        0.889
##    87     58       1        0    0.751  0.0595        0.643        0.877
##   104     60       1        0    0.739  0.0598        0.631        0.866
##   110     59       1        0    0.726  0.0601        0.618        0.854
##   129     57       1        1    0.714  0.0603        0.605        0.842
##   147     56       1        0    0.701  0.0606        0.592        0.830
##   194     55       1        0    0.688  0.0608        0.579        0.818
##   254     66       1        0    0.678  0.0608        0.568        0.808
```

Run a Cox Proportional Hazard regression using mother's age and infant's gender

```
# Cox Proportional Hazard regression
cox <- coxph(infants_surv ~ age + sex, data = infants)
summary(cox)
```

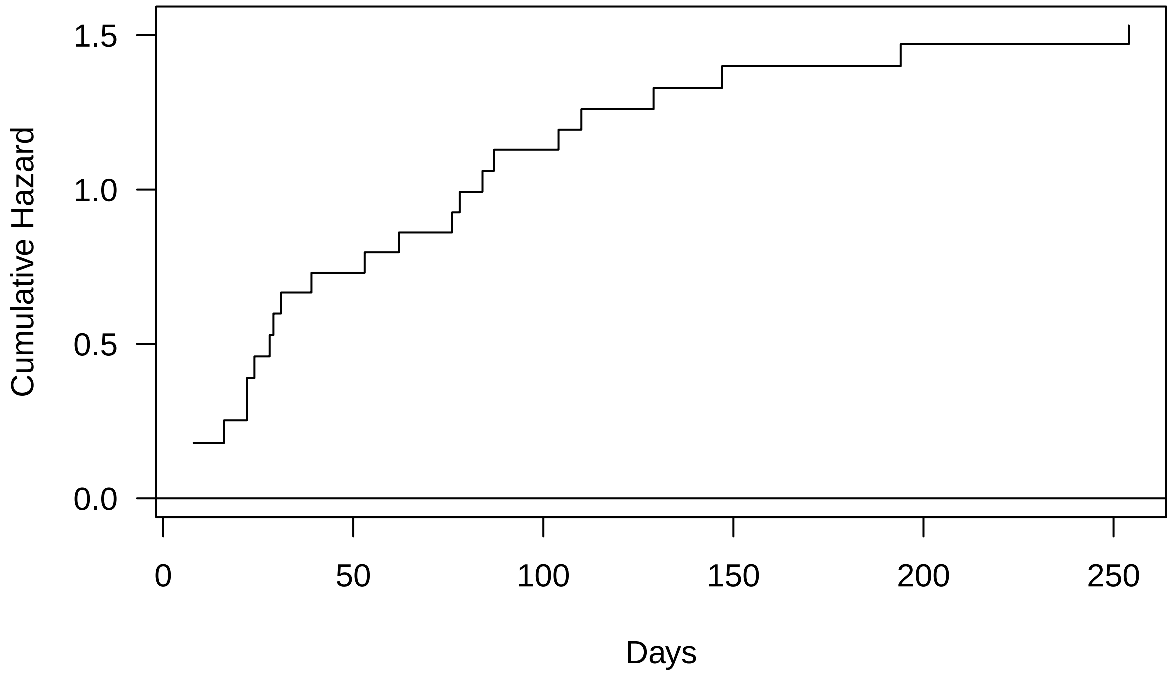
```
## Call:
## coxph(formula = infants_surv ~ age + sex, data = infants)
##
##   n= 105, number of events= 21
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## age    -0.04044   0.96037  0.04507 -0.897    0.370
## sexboy -0.48518   0.61559  0.44224 -1.097    0.273
##
##        exp(coef) exp(-coef) lower .95 upper .95
## age       0.9604      1.041    0.8792     1.049
## sexboy    0.6156      1.624    0.2587     1.465
##
## Concordance= 0.586  (se = 0.058 )
## Likelihood ratio test= 1.99  on 2 df,   p=0.4
## Wald test            = 2  on 2 df,   p=0.4
## Score (logrank) test = 2.03  on 2 df,   p=0.4
```

Interpret the Cox Proportional Hazard regression

- There is a 0.04 decrease in the expected log of the hazard for mother's age at infant's birth in 1 unit increase, holding sex constant.
- There is a 0.485 decrease in the expected log of the hazard for male babies compared to female, holding mother's age constant.
- Let's have a look at exponentiate parameter which estimates to obtain hazard ratios:
- Because for age, exp(-0.04) = 0.96, the hazard ratio of mother's age increase by 1 unit is 0.96. It indeicates that the babies are less likely to die when mother's age increase by 1 unit. We can understand it as there are 96 babies die when mother's age increase by 1 unit comparing with 100 babies die at mother's age. It means 1 unit age older of the mother, the babies deaths are 4% lower.
- Because for sexboy, exp(-0.485) = 0.62, the hazard ratio of male babies is 0.62 that of female babies. It indicates that male babies are less likely to die, as 62 male babies die for every 100 female babies. So male deaths are 38% lower than femal babies.

Plot the Cox Proportional Hazard model with Cumulative Hazard Function

```
cox_fit <- coxreg(infants_surv ~ age + sex, data = infants)
plot(cox_fit, xlab = "Days", ylab = "Cumulative Hazard")
```



##

Interpret the cumulative hazard: - Cumulative hazard function is supposed to have higher accuracy than the hazard function. We can see that our cumulative hazard function is a stair shape line with the rate under 1.5% during 250 days.The rate of change is flatter and slower at a certain constant lever after day 150, while the rate of change is sharper before day 150.

Make a Chisq test to assess the quality of this model

```
# Chisq test
drop1(cox, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## infants_surv ~ age + sex
##        Df    AIC     LRT Pr(>Chi)
## <none>    171.25
## age     1 170.12 0.87104   0.3507
## sex     1 170.42 1.16584   0.2803
```

- As the P-Values are 0.35 for mother's age and 0.28 for babie's gender, we can say it is not a statistically significant result.