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 Skellefte, 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.

Starting by loading the libraries.

```
1
2
3 # load libraries
4 pkgTest <- function(pkg){
5   new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
6   if (length(new.pkg))
7     install.packages(new.pkg,  dependencies = TRUE)
8   sapply(pkg,  require,  character.only = TRUE)
9 }
10
11 lapply(c("survival", "eha", "tidyverse", "ggfortify", "ggforest", "
     stargazer", "ggforestplot"),  pkgTest)
12 > # Fit a Cox proportional hazard model with mother's age and infant's
     gender as covariates
13 > fit <- coxph(Surv(exit, event) ~ m.age + sex, data = child)
```

```
> # Summarize the results
> summary(fit)
Call:
coxph(formula = Surv(exit, event) ~ m.age + sex, data = child)

  n= 26574, number of events= 5616

               coef exp(coef) se(coef)     z Pr(>|z|)
m.age       0.00762   1.00765  0.00213  3.58  0.00034 ***
sexfemale  -0.08221   0.92107  0.02674 -3.07  0.00211 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
            1

           exp(coef) exp(-coef) lower .95 upper .95
m.age          1.008      0.992     1.003     1.012
sexfemale      0.921      1.086     0.874     0.971

Concordance= 0.519  (se = 0.004 )
Likelihood ratio test= 22.5  on 2 df,   p=0.00001
Wald test            = 22.5  on 2 df,   p=0.00001
Score (logrank) test = 22.5  on 2 df,   p=0.00001

> # Create a summary table of the model
> stargazer(fit, type = "text")

================================================
                        Dependent variable:
                    ---------------------------
                                exit
------------------------------------------------
m.age                         0.008***
                               (0.002)

sexfemale                     -0.082***
                               (0.027)

------------------------------------------------
Observations                   26,574
R2                              0.001
Max. Possible R2                0.986
Log Likelihood                -56,503.500
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 coefficients for m.age and sexfemale indicate the hazard ratio for each unit increase in mother's age and for infants who are female, respectively. For example, the hazard ratio for m.age is 0.00762, which means that for each additional year of age, the hazard of child

mortality increases by a factor of 0.00762, holding all other variables constant.

The p-values for each covariate indicate whether or not they are statistically significant. In this case, both m.age and sexfemale are highly significant (p ¡ 0.001).

The concordance index (C-index) measures the predictive accuracy of the model, with values closer to 1 indicating better accuracy. In this case, the C-index is 0.519, which suggests that the model is only moderately accurate in predicting child mortality. For the variable m.age, the coefficient estimate is 0.008, which means that for a one-unit increase in m.age, the hazard rate increases by a factor of exp(0.008) = 1.008. The standard error for the estimate is 0.002, and the p-value is less than 0.01, indicating that this variable is statistically significant at the 0.01 level.

Similarly, for the variable sexfemale, the coefficient estimate is -0.082, which means that females have a hazard rate that is exp(-0.082) = 0.921 times that of males. The standard error for the estimate is 0.027, and the p-value is less than 0.01, indicating that this variable is also statistically significant.

The summary table also includes information about the goodness of fit of the model, including the number of observations, the R-squared value, the maximum possible R-squared value, and the results of various tests of model fit, such as the Wald test, LR test, and Score (Logrank) test. The p-values for these tests indicate whether the model fits the data significantly better than a null model with no predictors.

```
> # Create a dataframe with the coefficient and confidence intervals
> coef_df <- data.frame(
+   Variable = c("Mother's Age", "Infant's Gender (Female)"),
+   Estimate = c(0.008, -0.082),
+   Lower_CI = c(0.004, -0.135),
+   Upper_CI = c(0.012, -0.029)
+ )
> coef_df
                  Variable Estimate Lower_CI Upper_CI
1             Mother's Age    0.008    0.004    0.012
2 Infant's Gender (Female)   -0.082   -0.135   -0.029
```