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.

First I loaded the dataset into R:

```
library("eha")
data(infants)
```

Then I loaded the packages needed to fit the Cox Proportional Hazard Model:

```
install.packages(c("survival", "survminer"))
library("survival")
library("survminer")
```

Then, I ran a cox regression to see how both variables indiviudally affected survival. I did this with the following code:

```
## Apply univariate coxph function to multiple covariates at once (age and sex
    )
covariates <- c("age", "sex")
univ_formulas <- sapply(covariates,
```

```
                        function(x) as.formula(paste('Surv(exit, event) ~', x)
    ))

univ_models <- lapply(univ_formulas, function(x) {coxph(x, data = infants)})

## Extract data
univ_results <- lapply(univ_models,
                       function(x) {
                         x <- summary(x)
                         p.value <- signif(x$wald["pvalue"], digits = 2)
                         wald.test <- signif(x$wald["test"], digits = 2)
                         beta <- signif(x$coef[1], digits = 2); # coefficient
    beta
                         HR <- signif(x$coef[2], digits = 2); # exp beta
                         HR.confint.lower <- signif(x$conf.int[, "lower .95"],
     2)
                         HR.confint.upper <- signif(x$conf.int[,"upper .95"
    ],2)
                         HR <- paste0(HR, " (",
                                      HR.confint.lower, "-", HR.confint.upper,
     ")")
                         res<-c(beta, HR, wald.test, p.value)
                         names(res)<-c("beta", "HR (95% CI for HR)", "wald.
    test",
                                       "p.value")
                         return(res)
                         #return(exp(cbind(coef(x),confint(x))))
                       })

res <- t(as.data.frame(univ_results, check.names = FALSE))
```

This resulted in the following output:

```
as.data.frame(res)
```

```
     beta HR (95% CI for HR) wald.test p.value
age -0.035    0.97 (0.87-1.1)      0.47    0.49
sex  -0.51     0.6 (0.25-1.4)       1.3    0.25
```

The interpretation of this output is as follows: The output shows regression beta coefficients ("beta"), the effect sizes (HRatios), statistical significance of each variable in relation to overall survival (pvals)

Neither variable (age of mother or sex) have statisticalluy significant coefficients as the p.values are .49 and .25 respectively. Both of these pvalues are well above common significance levels (of say .1, .05 etc). Both age and sex have negative beta coefficients. This implies that having an older mother and being a boy is associated with better survival, however it is important to note that still, these coefficients are not statistically significant so we cannot say the difference between groups is differentiable from zero really.

NEXT, I interpret how the two covariates affect survival jointly. I do this with multivariate cox regression:

```r
# COX regression of exit (age at death) on time constant covariates:

res.cox <- coxph(Surv(exit, event) ~ age + sex, data = infants)
summary(res.cox)
```

Which outputs the following summary:

```
Call:coxph(formula = Surv(exit, event) ~ age + sex, data = infants)
  n= 105, number of events= 21
              coef exp(coef) se(coef)      z
age    -0.03021   0.97024  0.04957 -0.610
sexboy -0.49007   0.61258  0.44259 -1.107

        Pr(>|z|)
age        0.542
sexboy     0.268

       exp(coef) exp(-coef) lower .95 upper .95
age       0.9702      1.031    0.8804     1.069
sexboy    0.6126      1.632    0.2573     1.458

Concordance= 0.591  (se = 0.058 )
Likelihood ratio test= 1.69  on 2 df,   p=0.4
Wald test            = 1.72  on 2 df,   p=0.4
Score (logrank) test = 1.76  on 2 df,   p=0.4
```

The interpretation of these results is as follows: The p values again for the 3 tests run are not significant. This indicates that the model is not significant, and we cannot reject the null hypothesis that all of the betas are 0. In the multivariate analysis, the covariates are ot significant (not surprising), they weren't significant in previous model either. If the p vlaues were significant, the fact that the Hazard Ratio (exp(coef)) for age is .97, and for sex is .61 could be interpreted as follows: Holding other covariates constant, an infant that is a boy reduces the hazard by a factor of .61. Being a boy holding all else constant would reduce the risk of death. Holding the other covariates constant, an infant with an older mother is also less likely to die. HOWEVER, the CONFIDENCE INTERVALS for both hazard ratios include 1. This alongside the non-significant p values for both indicates that age of mother and sex of infant make relatively small (or null) contributions to likelihood of survival.