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.

1 Theory behind Cox Proportional Hazard Model

The Cox proportional Hazard model is a survival model used to model the relationship between the time to an event and predictor variables. It relates the time that passes before an event occurs to covariates that may be associated with that quantity of time.

The basic equation of the model is as follows:

$$h(t) = h0(t) * e^{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p} \tag{1}$$

Where h(t) is the expected hazard at time t, $h_0(t)$ is the baseline hazard, expected when all the predictors X_1, X_2, X_p and all the covariates between them are equal to zero.

If we divide both sides of the equation by $h_0(t)$, we get the following equation:

$$\frac{h(t)}{h_0(t)} = e^{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p} \tag{2}$$

If we take a natural log of both sides of this equation, we get the following:

$$ln(\frac{h(t)}{h_0(t)}) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{3}$$

The β_1 value in this case represents the change in the expected log of the hazard ratio relative to a one unit change in X_1, holding all other predictors constant. The β coefficients in this case are what our results table will give us, after we run the regression.

2 Running the regression in R

1. We begin by importing the libraries we will need to generate the output.

```
# 1. Load in libraries
library(eha)
library(survival)
library(stargazer)
```

2. We then load in the "child" dataset.

```
# 2. Load in data
child <- child
```

3. We then create the survival object from the data.

```
# 3. Creating the survival object.
child_surv <- with(child, Surv(enter, exit, event))
```

4. We then run the Cox Proportional Hazard Regression on the data, with mother's age and infant's gender as covariates.

```
# 4. Run a Cox Proportional Hazard Regression on the
# data with mother's age and infant's gender as
# covariates.

cox.reg <- coxph(child_surv ~ m.age + sex,
data = child)
```

5. We use stargazer to generate output for us, as follows:

```
stargazer(cox.reg,
type = "latex")
```

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

3 Interpreting our regression results

The results table indicates to us that both predictor variables are statistically significant at the 1% percent level, for a 0.01 alpha level of significance ($\alpha = 0.01$) This is to say that both predictors have a significant non-zero relationship with the outcome variable.

The coefficient estimate for `m.age` is 0.008. This means that for a one year increase in the age in the mother, we would expect a 0.008 increase in the log of the hazard ratio.

```
exp(0.008)
```

When we calculate `exp(0.008)`, we get a result of 1.008032. This means that for a one unit increase in the age of the mother, we would expect the hazard ratio (HR) of the death for the child to increase by 0.8%, holding all other predictors constant.

The coefficient estimate for `sexfemale` is -0.082. This means that there is a 0.082 decrease in the expected log of hazard for female infants compared with male ones, holding all other predictors constant.

```
exp(-0.082)
```

When we calculate `exp(-0.082)`, we get a result of 0.921272. This means that the hazard ratio for female children is 8% lower than for male children, meaning that around 92 female infants die for every 100 male infants.

The results of the Wald, Likelihood-Ratio and Score (Logrank) test, with statistically significant results at the $\alpha = 0.01$ level, suggest that the model is a good fit for the data.

Our R^2 value indicates that the predictor variables only explain a small proportion, namely 0.1% of the variability in the outcome variable child survival.

We can represent our coefficients as follows:

$$ln(\frac{h(t)}{h_0(t)}) = 0.008 * (M.AGE) - 0.082 * (SEX.FEMALE) \tag{4}$$

Or as follows:

$$\frac{h(t)}{h_0(t)} = e^{0.008 * M.AGE - 0.082 * SEX.FEMALE} \tag{5}$$