# Problem Set 4

Applied Stats II

Due: April 4, 2022

## 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 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.**

To preform a Cox Proportional Hazard model, firstly the data was explored, shown below using str(infants):

```
'data.frame': 105 obs. of  11 variables:
 $ stratum: int  1 1 1 2 2 2 3 3 3 4 ...
 $ enter  : int  55 55 55 13 13 13 361 361 361 2 ...
 $ exit   : int  365 365 365 76 365 365 365 365 365 16 ...
 $ event  : int  0 0 0 1 0 0 0 0 0 1 ...
 $ mother : Factor w/ 2 levels "alive","dead": 2 1 1 2 1 1 2 1 1 2 ...
 $ age    : int  26 26 26 23 23 23 24 24 24 28 ...
 $ sex    : Factor w/ 2 levels "girl","boy": 2 2 2 1 1 1 2 2 2 1 ...
 $ parish : Factor w/ 2 levels "other","Nedertornea": 2 2 2 2 2 2 2 2 2 2 ...
 $ civst  : Factor w/ 2 levels "married","unmarried": 1 1 1 1 1 1 1 1 1 1 ...
```

```
$ ses    : Factor w/ 2 levels "other","farmer": 2 2 2 1 1 1 1 1 1 1 ...
$ year   : num  1877 1870 1882 1847 1847 ...
```

I focused on the data necessary for the model: age, sex, enter, exit, event. Age, enter, exit and event are encoded as integers and sex is a factor, so this all looks good so we can continue:

Next, visualisation was preformed using the following code:

```
infants_surv <- with(infants, Surv(enter, exit, event))
km <- survfit(infants_surv ~ 1, data = infants)
autoplot(km)
```

This produced the following:

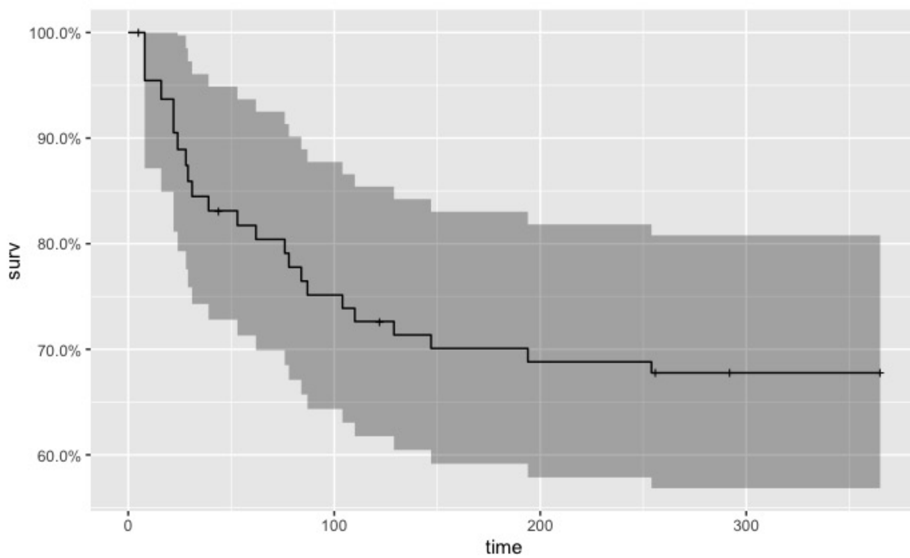


Figure 1: Kaplan Meier Plot

The above Kaplan Meier plot is a descriptive way to visualise the Data. It starts at 0 when 100% of the sample were alive and then shows the decline in survival overtime. It continues until the first event, death in this case, and then it declines horizontally for each event. Furthermore, the plot continues until the curve hits zero or the study ends, and in this case it was when study ended at about a year( 360 days).

Next, I explored the covariate "sex" further, again, using a Kalplan Meier Plot.

```
km_sex <- survfit(infants_surv ~ sex, data = infants)
autoplot(km_sex)
```

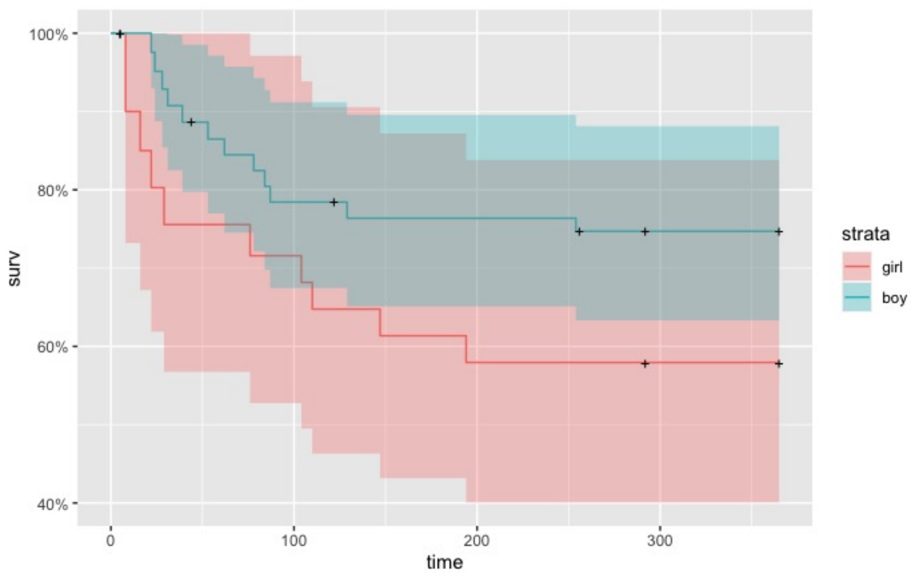


Figure 2: Kaplan Meier Plot, with covariate "sex"

From this graph it visualises that the survival rate of female infants is about 20% lower than male infants, a Cox Proportional Hazard model will be carried out below to further understand this relationship.
As sex was a factor this was helpful to visualise, however I did not plot age, as the numerical information would not be as easy to comprehend on a km plot. To further understand age, I moved onto the Cox proportional Hazard Model.

**Fit a Cox Proportional Hazard model using mother's age and infant's gender as covariates**

Using the Surv function for the outcome variable, the following code was used to preform the Cox Proportional Hazard model:

```
add_surv <- coxph(Surv(enter, exit, event) ~ age + sex, data = infants)
```

Below is the output of the Cox Proportional Hazard model:

Table 1:

| | *Dependent variable:* |
|---|---|
| | enter |
| age | −0.040 |
| | (0.045) |
| sexboy | −0.485 |
| | (0.442) |
| Observations | 105 |
| $R^2$ | 0.019 |
| Max. Possible $R^2$ | 0.800 |
| Log Likelihood | −83.626 |
| Wald Test | 2.000 (df = 2) |
| LR Test | 1.992 (df = 2) |
| Score (Logrank) Test | 2.034 (df = 2) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Interpretation**

*Sex*

There is a 0.485 decrease in the expected log odds of the hazard for male babies compared to female, holding mothers age constant. This can also be interpreted as a multiplicative effect of exp(-0.485) = 0.6156972 for male babies, when holding mothers age constant. This portrays what was shown in the Kaplan Meier plot above, whereby due to a decrease in the male hazard rate, the survival rate of females is lower than that of males.

*Mothers age*

For a one unit increase in age, there is a 0.040 decrease in the expected log odds of the hazard for infants, holding sex constant. This can also be interpreted as a multiplicative effect of exp(-0.040) = 0.9607894 on the Hazard rate for infants, for every one unit increase in mothers age, while holding sex constant. Suggesting that the higher the mothers age, the lower the Hazard rate.