# Problem Set 4

Applied Stats II

Due: April 16, 2023

## 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.

```

# Load data
data(child)
head(child)

# Build a survival object out of the 'child' data.frame
child_surv <- with(child, Surv(enter, exit, event))

# Run a Cox Proportional Hazard regression
cox <- coxph(child_surv ~ m.age + sex, data = child)
summary(cox)
drop1(cox, test = "Chisq")
stargazer(cox, type = "text")

```

We can plot this survival function using a Kaplan Meyer plot:

```
# Plot
km <- survfit(child_surv ~ 1, data = child)
summary(km, times = seq(0, 15, 1))
plot(km, main = "Kaplan-Meier Plot", xlab = "Years", ylim = c(0.7, 1))
```

```
# Run a Cox Proportional Hazard regression
cox <- coxph(child_surv ~ m.age + sex, data = child)
summary(cox)
drop1(cox, test = "Chisq")
stargazer(cox, type = "text")
```

Please view output on next page:

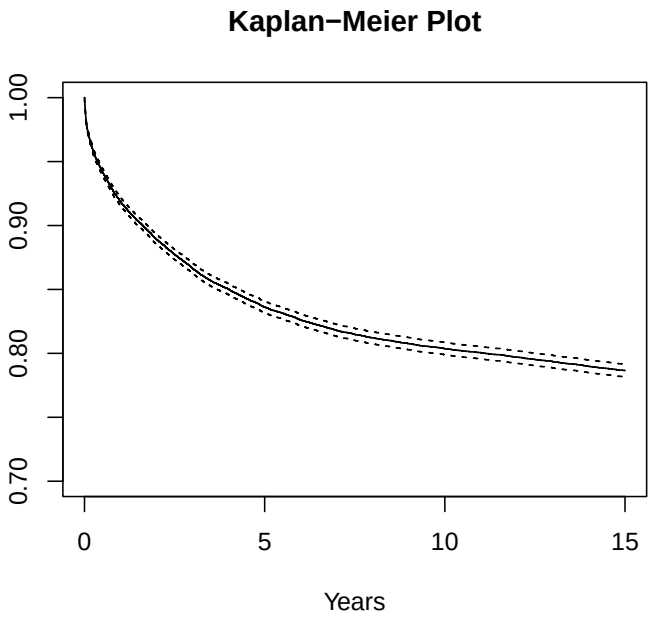


Figure 1: Child Survival Kaplan Meier plot

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 |

**Analysis:** The results show that both mother's age and infant's gender are statistically significant predictors of child mortality, with p-values less than 0.01.

For mother's age, the coefficient estimate of 0.008 indicates that for each one-year increase in mother's age, the hazard of child mortality increases by a factor of 1.008 (exp(0.008)).

For infant's gender, the coefficient estimate of -0.082 indicates that the hazard of mortality is lower for female infants than for male infants by a factor of 0.921 (exp(-0.082)).

The drop1() function allows us to asses the model quality by removing each predictor one at a time from the model and comparing the fit of the reduced model to the original model using an LRT.

```
> drop1(cox, test = "Chisq")
Single term deletions

Model:
child_surv ~ m.age + sex
Df    AIC     LRT  Pr(>Chi)
<none>    113011
m.age   1 113022 12.7946 0.0003476 ***
sex     1 113018  9.4646 0.0020947 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here we can see that that removing either mother's age or infant's gender from the model leads to a significant deterioration in the model fit based on the likelihood ratio test. This indicates that both predictors are important in predicting child survival.