Problem Set 4

Applied Stats II

Due: April 12, 2024

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 Friday April 12, 2024. 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.

I load and inspect my data:

```
# load data
data <- child

# Inspect data
str(data) # Call structure of the data
summary(data) # Call summary statistics for each column

head(data, 5) # Call first 5 observations
tail(data, 5) # Call final 5 observations
colnames(data) # Call column names
lapply(data, typeof) # Call datatypes for each column
# View(data) # View whole dataset
print(unique(data$socBranch)) # Call names of the father's working branch
```

I create the survival object `child_surv` and, to further explore the data, plot the Kaplan-Meier curves for Overall Survival and across the two variables for Mother's Age `m.age` and Child's Sex `sex`, as shown in Figures 1-3.

Note that on average, the expected survival probability for a child falls dramatically in their first few years of life, with the rate of decline gradually reducing up to age 15, at which point our data ends. Note the relatively tight margins along the KM curve, indicating relative stability in the estimated survivability of children over time.

```
# Create survival object
child_surv <- with(data, Surv(enter, exit, event))
class(child_surv) # confirms survival object class
```

```
# Overall Survival For reference
pdf("overall.pdf", height=8.5, width=11)
km <- survfit(child_surv ~ 1, data = child)
summary(km, times = seq(0, 15, 1))
autoplot(km, #main = "Kaplan-Meier Plot: Mother's Age",
         xlab = "Years",
         ylab = "Estimated Survival Probability",
         ylim = c(0.7, 1))
dev.off()
```

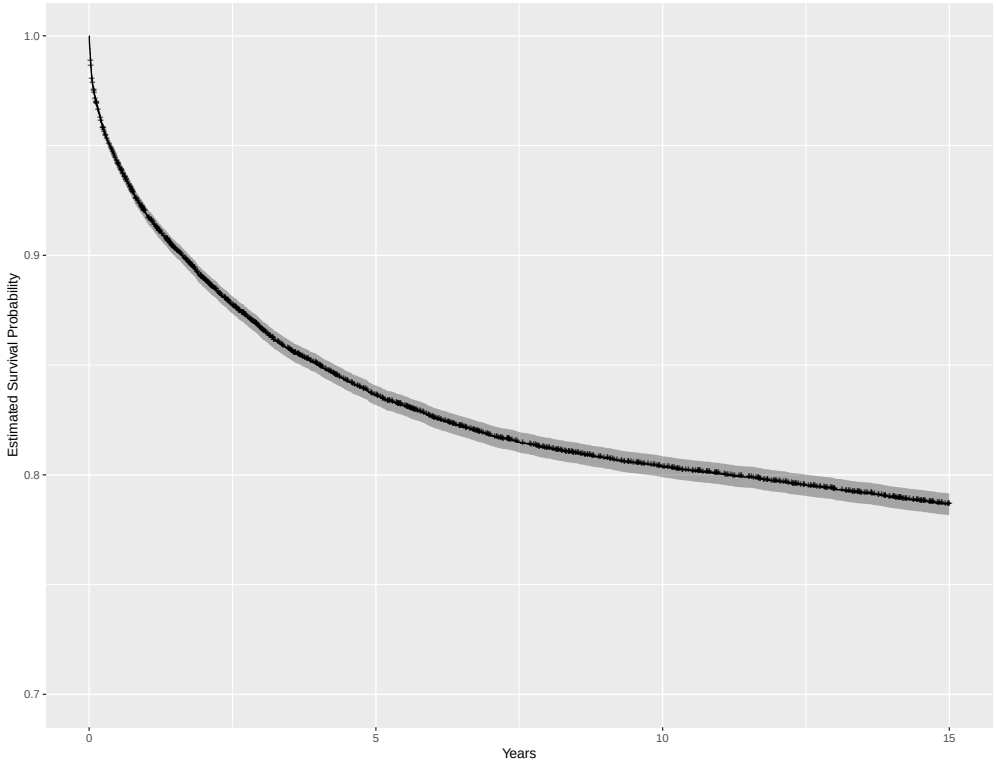

Figure 1: Kaplan-Meier Plot: Overall Survival Over Time

Note that Mother's Age is a continuous variable, and so for the purposes of plotting the KM curve, which requires a categorical predictor variable, I create a new object with variable `m.age_bins`, which separates ages across five categories of approx. 10-years each (15-17, 18-24, 25-34, 35-44, and 45-51). Note the wide variation for the 15-17yrs - this is likely due to including a very small number of observations compared to the other age bands (see Table 1). That said, additional factors may affect the survivability outcomes for the children of especially young mothers, such as family connections, financial stability/socio-economic background of mother's parents, and the general health of a younger mother undergoing pregnancy and associated medical complications.

```
# Create second dataset for plotting Mother's Age
data1 <- data
# Find max/mean age
max(data1$m.age) # is 50.846
min(data1$m.age) # is 15.828
# Define the boundaries for the age group
age_breaks <- c(15, 18, 25, 35, 45, 51)
# Create labels for the age groups
age_labels <- c("15-17", "18-24", "25-34", "35-44", "45-51")
# Create the new variable m.age_bins by categorizing age into the specified
    groups
data1$m.age_bins <- cut(data$m.age, breaks = age_breaks, labels = age_labels,
    right = FALSE)
# Inspect new variable and check level order is correct
print(data1$m.age_bins)
table(data1$m.age_bins) # Check contingency table
levels(data1$m.age_bins) # Check order is correct

# Create second survival object
child_surv1 <- with(data1, Surv(enter, exit, event))
class(child_surv) # confirms survival object class
```

Excluding the 15-17yrs age group, on average, the mother's age appears to have a distinct effect on the estimated survival probability of the child, where children from younger mothers tend to have higher survival across all lived years, controlling for all other variables (though there is minor overlap in variation across lived years of the child).

```
# Create contingency table for Mother's Age bins
table <- as.data.frame(table(data1$m.age_bins))
xtable(table, type = "latex")
```

```
# Survival by Mother's Age
pdf("age.pdf", height=8.5, width=11)
km_age <- survfit(child_surv1 ~ m.age_bins, data = data1)
summary(km_age, times = seq(0, 15, 1))
autoplot(km_age,
         #main = "Kaplan-Meier Plot: Mother's Age",
         xlab = "Years",
         ylab = "Estimated Survival Probability")
dev.off()
```

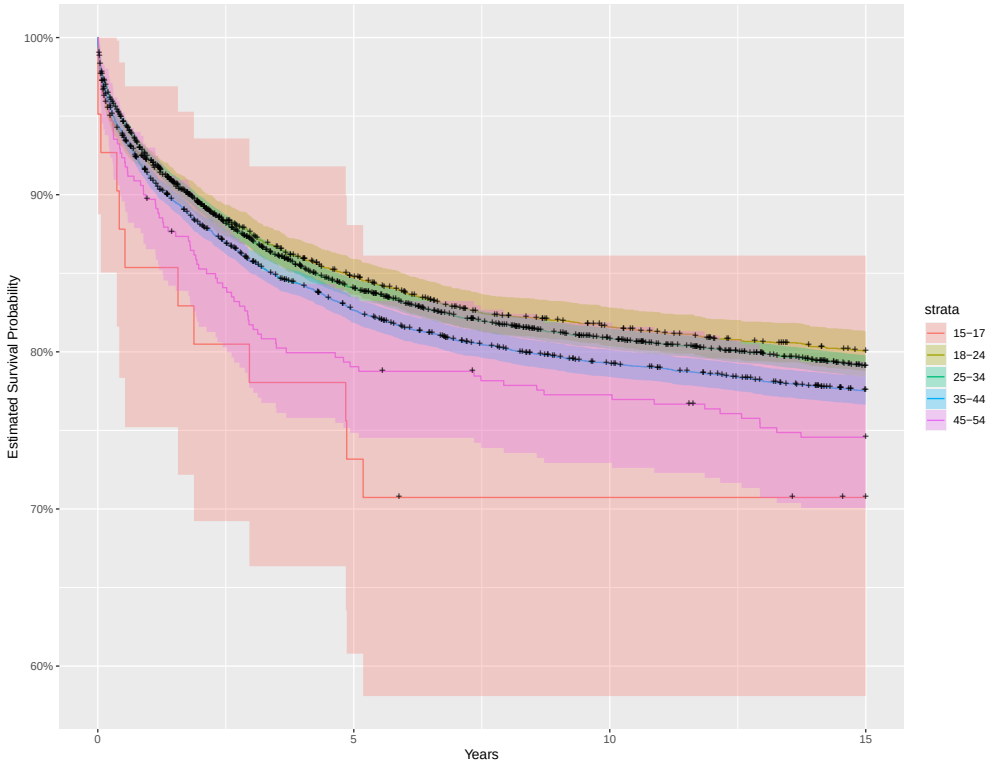

Figure 2: Kaplan-Meier Plot: Mother's Age

Age Group	Freq
15-17	41
18-24	3876
25-34	13805
35-44	8512
45-54	340

Table 1: Number of Mothers in Each Age Group

Note that Child's Sex appears to have a distinct effect on the (estimated) survival probability for the child - on average, females appear to have a higher probability of survival compared to males from birth and across all lived years, controlling for all other variables.

```
# Survival By Infants Sex
pdf("sex.pdf", height=8.5, width=11)
km_sex <- survfit(child_surv ~ sex, data = child)
summary(km_sex)
autoplot(km_sex, #main = "Kaplan-Meier Plot: Child's Sex",
         xlab = "Years",
         ylab = "Estimated Survival Probability",
         ylim = c(0.7, 1))
dev.off()
```

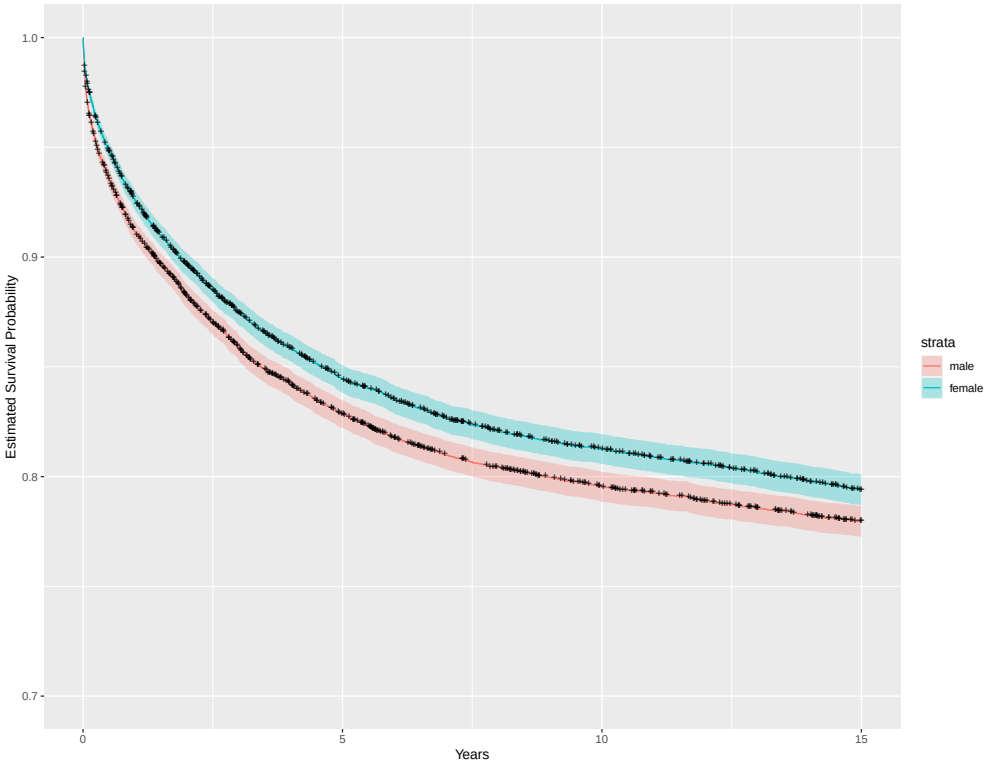

Figure 3: Kaplan-Meier Plot: Child's Sex

Next, I fit a Cox Proportional Hazard model using mother's age and infant's sex as covariates. I create two models - one additive and one interactive, to further explore the relationship between the two predictors and the survival probability of a child. Comparing both, we see the average partial effects on Survival Probability by Mother's Age and Child's Sex are statistically significant for the Additive model, while only Mother's Age is statistically significant in the Interaction model. They otherwise have the same or similar values in other notable figures on the table, and so I will continue with the additive model for further analysis.

Regarding interpreting Table 2:

- On average, for every 1-unit increase in the Mother's Age, the hazard (risk) of child survival increases by 0.008 units, holding all other variables constant. This estimate is statistically significant. In other words, on average, for every increase in 1 year in the Mother's Age, the child's probability of survivial decreases by 0.008 percentage points, holding all other variables constant.

- On average, a child who is female has a decreased hazard to their survival by 0.082 units, holding all other variables contant. This estimate is statistically significant. In other words, on average, children assigned female at birth have an average increase in 0.082 percentage points in comparison to children assigned male at birth, holding all other variables constant.

- The R^2 is 0.001, indicating that within the given data, the variations in Mother's Age and a Child's Sex co-vary with almost none of the variation in a child's estimated Survival Probability.

- The Wald Test is statistically significant, which gives evidence to reject the null hypothesis that no predictors included in the model have an effect on the outcome variable, and gives evidence for the alternative hypothesis that at least one of the included predictors - Mother's Age and/or Child's Sex - has an effect on a child's Estimated Survival Probability.

- The Score (Logrank) Test is statistically significant, which gives evidence to reject the null hypothesis that just the null model (mean) is better than the chosen model for estimating Survivability Probability for a child, and gives evidence for the alternative hypothesis that the chosen model is a better fit to the data on average.

In conclusion, while there is significant evidence supporting the validity of this model, due to the oerall poor model fit of The $R^2 = 0.001$, another model or alternative explanatory variables may be more appropriate or useful for estimating the Survivability Probability of children born in Skellefteå, Sweden from 1850 to 1884.

```
# Create CPH models
cox1 <- coxph(child_surv ~ m.age + sex, data = data)
summary(cox1)
```

```r
stargazer(cox1, type = "latex")

cox2 <- coxph(child_surv ~ m.age * sex, data = data)
summary(cox2)
stargazer(cox2, type = "latex")
```

Table 2: Cox Proportional Hazard Model: Additive

	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

Table 3: Cox Proportional Hazard Model: Interaction

	Dependent variable:
	child_surv
m.age	0.007^{**}
	(0.003)
sexfemale	-0.127
	(0.140)
m.age:sexfemale	0.001
	(0.004)
Observations	26,574
R^2	0.001
Max. Possible R^2	0.986
Log Likelihood	$-56{,}503.430$
Wald Test	22.530^{***} (df = 3)
LR Test	22.624^{***} (df = 3)
Score (Logrank) Test	22.562^{***} (df = 3)
Note:	*p<0.1; **p<0.05; ***p<0.01