Problem Set 4

Jack Merriman

Submitted: April 6, 2023

Question 1

```r
#create a survival object
survival <- Surv(data$enter, data$exit, data$event)
#specify the Cox Proportional Hazard model
mod1 <- coxph(survival ~ m.age + sex, data = data)
stargazer(mod1)
```

Table 1:

	Dependent variable:
	survival
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

Our coefficients are statistically significant at $p < 0.01$ as shown by the triple asterisks in the stargazer plot, we will use `anova` against a null model to conduct a χ^2 test on the model:

```r
#chisq test
modnull <- coxph(survival ~ 1, data = data)
stargazer(anova(modnull, mod1))
```

Table 2:

Statistic	N	Mean	St. Dev.	Min	Max
loglik	2	−56,509.110	7.961	−56,514.740	−56,503.480
Chisq	1	22.518		22.518	22.518
Df	1	2.000		2	2
Pr(>|Chi|)	1	0.00001		0.00001	0.00001

We can see the p-value is infinitesimally small, and therefore our model is a good predictor of survival.

With our model and its coefficients statistically significant, we can interpret the coefficients with as such:

`m.age`: Holding sex constant, an increase of one year in the age of the child's mother is associated with a 0.008 increase in the expected log of the hazard rate for the child.

`sexfemale`: Holding the age of the child's mother constant, girls have an average decrease of 0.082 in their expected log of the hazard rate relative to boys.

We can further our interpretations by exponentiating the coefficients to examine the hazard ratios themselves:

```r
#interpret coefficients through exponentiation
stargazer(exp(mod1$coefficients))
```

Table 3:

m.age	sexfemale
1.008	0.921

`m.age`: Holding sex constant, an increase of one year in the age of the child's mother is associated with a 0.8% increase in the likelihood of the child dying.

`sexfemale`: Holding the age of the child's mother constant, girls are 8% less likely to die than boys.

Although there is no theoretical reason to believe that an interaction effect would increase the predictive power of the model, we will create a second interactive model to test:

```r
#specify interactive model
mod2 <- coxph(survival ~ m.age * sex, data = data)
stargazer(mod2)
```

Table 4:

	Dependent variable:
	survival
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

We observe that the interaction term in the model (`m.age:sexfemale`) is not statistically significant, and that the sex term has lost its statistical significance. We will conduct another χ^2 test to see if the model as a whole is more predictive, to be sure.

```r
#chisq test
stargazer(anova(mod1, mod2))
```

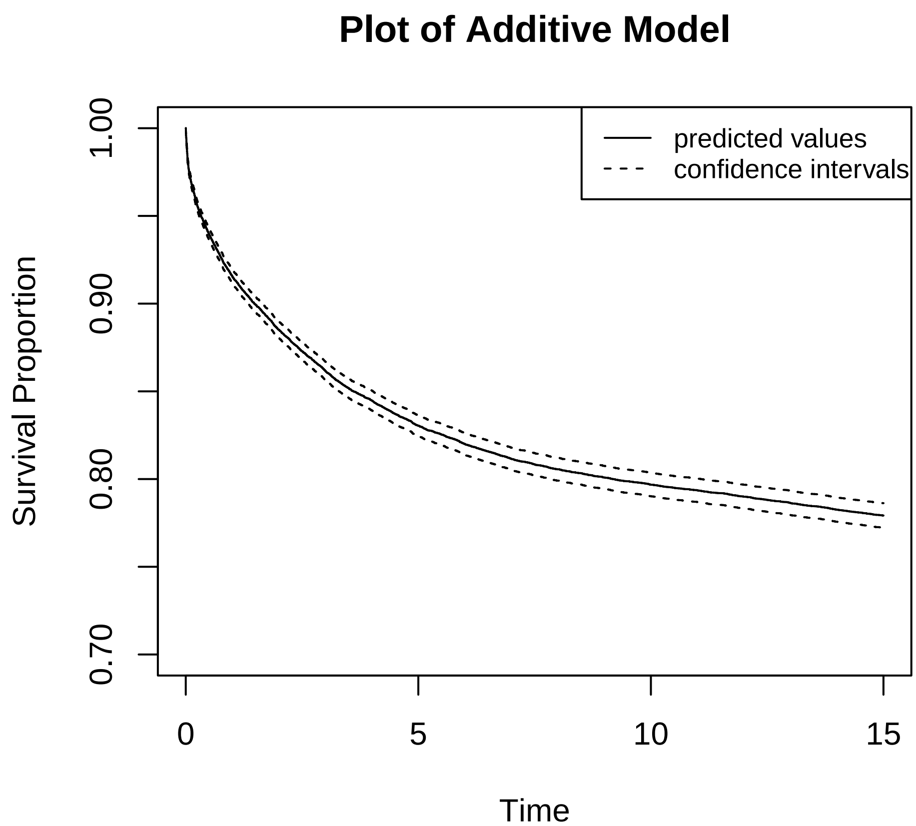
Table 5:

Statistic	N	Mean	St. Dev.	Min	Max
loglik	2	−56,503.460	0.038	−56,503.480	−56,503.430
Chisq	1	0.106		0.106	0.106
Df	1	1.000		1	1
Pr(>|Chi|)	1	0.744		0.744	0.744

With a p-value of 0.74, the model does not pass any significant critical threshold, so we can reject it being more predictive than the existing model which also has significant coefficients.

We are satisfied with the additive predictive power of the mother's age and the child's sex on their hazard ratio. Below are plots illustrating predicted values of the model:

```r
#create a fitted values object for plotting
modfit <- survfit(mod1)
plot(modfit, ylim = c(0.7,1), xlab = 'Time', ylab = 'Survival Proportion',
     main = 'Plot of Additive Model')
legend('topright',legend = c('predicted values', 'confidence intervals'),
       lty = c(1,2), cex = 0.8)
```



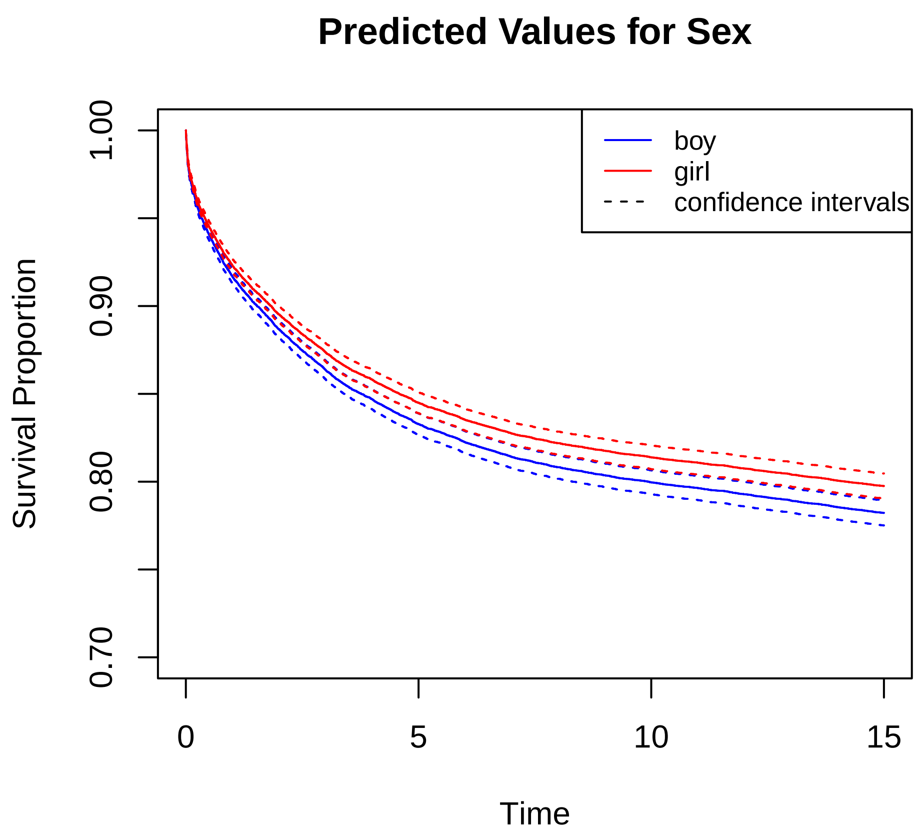
```r
#Find the modal age to set a base value for the predicted sex data
table(round(data$m.age))
```

```
  16   17   18   19   20   21   22   23   24   25   26   27   28   29   30 
   5   12   66  142  260  438  606  846  999 1143 1250 1302 1400 1444 1490 

  31   32   33   34   35   36   37   38   39   40   41   42   43   44   45 
1434 1435 1374 1439 1259 1285 1139 1122  997  859  818  638  507  377  247 

  46   47   48   49   51 
 132   66   32    9    2 
```

```r
#modal age is 30 with 1490 observations
predat1 <- with(child, data.frame(sex = c('male', 'female'), m.age = 30.0))
plot(survfit(mod1, newdata = predat1), conf.int = TRUE, ylim = c(0.7,1),
     col = c('blue', 'red'), xlab = 'Time', ylab = 'Survival Proportion',
     main = 'Predicted Values for Sex')
legend('topright',legend = c('boy', 'girl', 'confidence intervals'),
        col = c('blue', 'red', 'black'), lty = c(1,1,2), cex = 0.8)
```



```r
#Fit predicted age data to ten either side of the modal age
predat2 <- with(child, data.frame(sex = 'male', m.age = c(20.0, 30.0, 40.0)))
plot(survfit(mod1, newdata = predat2), conf.int = TRUE, ylim = c(0.7,1),
     col = c('darkgreen', 'orange', 'purple'), xlab = 'Time', ylab = 'Survival Proportion',
     main = 'Predicted Values for Age')
legend('topright',legend = c('20 year-old mother', '30 year-old mother',
                              '40 year-old mother', 'confidence intervals'),
       col = c('darkgreen', 'orange', 'purple', 'black'),
       lty = c(1,1,1,2), cex = 0.8)
```