MATH9952 Modern Applied Statistical Models
ASSIGNMENT MARKING SCHEME

Student No. D16126734 Name: JERRY KIELY

Assignment 5: SURVIVAL ANALYSIS

Part 1:	10
Part 2:	30
Part 3:	36
Total Marks from 100	76

Comments:

A basic but solid report for the Cox PH model.

It would have benefitted from more detailed analysis of the results - Perhaps present Hazard ratios and their respective CI's?

What about comparing the different BMI groups using customised hypotheses?

The Worcester Heart Attack Study

Jerry Kiely

19 April 2017

Introduction

The main goal of this study is to describe factors associated survival rates following hospital admission for acute myocardial infarction (MI), i.e. heart attack.

Name	Description	Codes / Values
id	Identification Number	
age	Age at Hospital Admission	Years
gender	Gender	0 = Male, 1 = Female
hr	Initial Heart Rate	Beats per minute
bmi_group	Body Mass Index category	low, medium, high, very.high
cvd	History of Cardiovascular Disease	0 = No, 1 = Yes
chf	Congestive Heart Complications	0 = No, 1 = Yes
miord	MI Order	0 = First, 1 = Recurrent
los	Length of Hospital Stay	Days from Hospital Admission
lenfol	Time variable	Days from Hospital Admission Date to Date of Last Follow-up
fstat	Event indicator	0 = Alive (censored), 1 = Dead

The Data

Before fitting a model to the data let's have a look at the behaviour of some of the predictors in isolation, beginning with Gender. From Figure 1 we can see that both the survival functions are reasonably close, with the female group being slightly worse than the male.

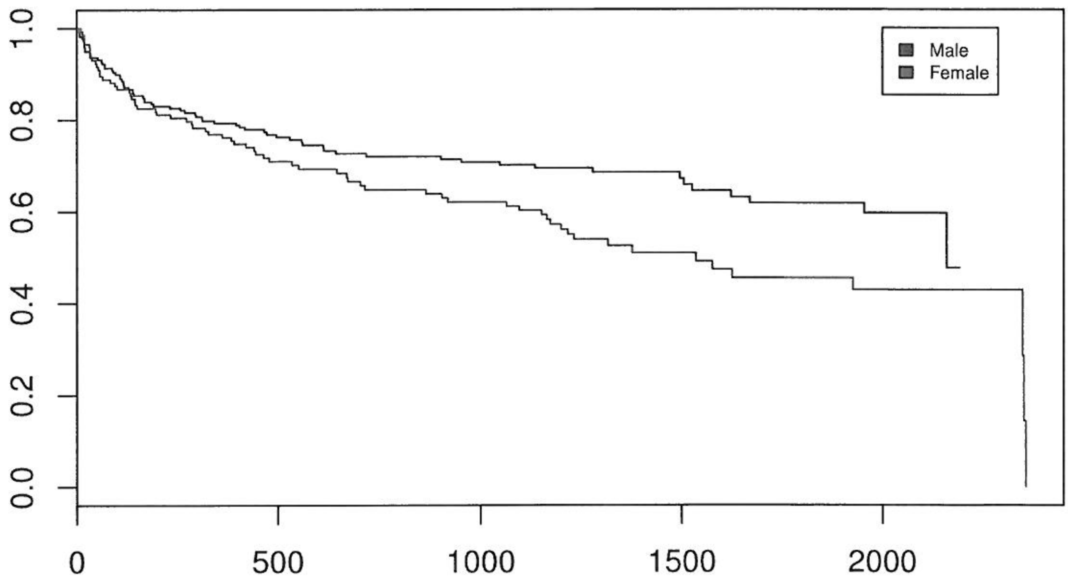


Figure 1: Survival Function by Gender

Now lets look at BMI Group. As can be seen from Figure 2, there is a lot of variation in the survival functions for the different BMI groups. What's interesting is that the low and very high groups seem to have similar survival functions, and the medium BMI group seems to have the worst survival function.

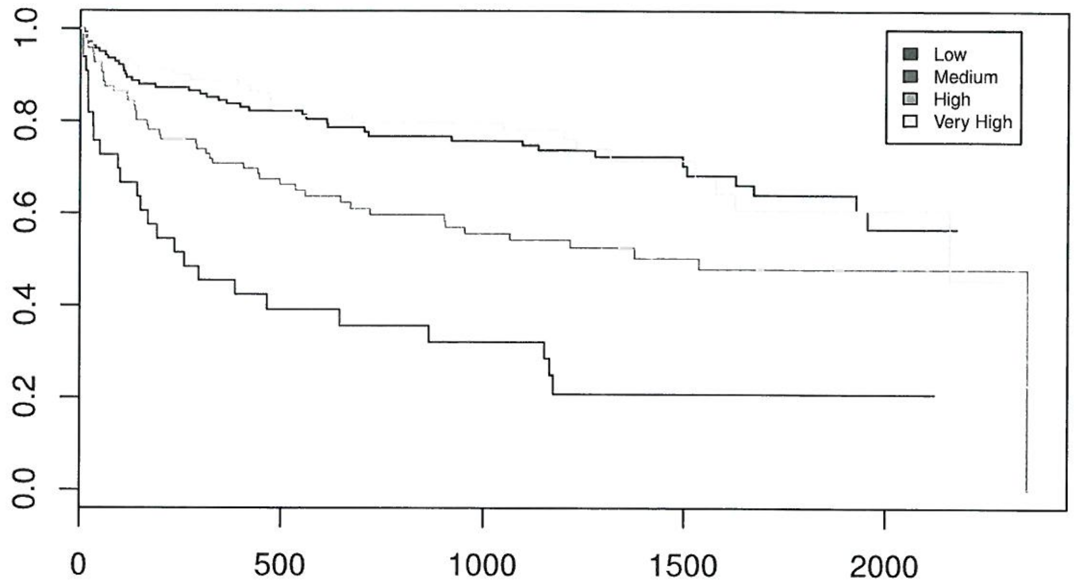


Figure 2: Survival Function by BMI Group

Both these predictors look to be of interest, and there does seem to be significant difference between the survival functions for each group within these predictors. Lets move on to fitting a full model and seeing which predictors are significant and which are not.

The Model

To begin with we fit a Cox PH model with all predictors and look to see which, if any, predictors can be pruned as insignificant. As can be seen from Table 2, the predictors with an AIC value less than the intercept's AIC value are cvd, los, and miord. Also, by looking at the p-values (right-most column) we see that these predictors have values which would also indicate insignificance.

? no interest in a Coxph Model – do you mean ~~a~~ "null" model?

Table 2: table detailing statistics for all predictors

	Df	AIC	LRT	Pr(>Chi)
	NA	1350.595	NA	NA
age	1	1403.761	55.1661583	0.0000000
factor(gender)	1	1353.165	4.5696690	0.0325429
hr	1	1356.081	7.4853886	0.0062202
factor(bmi_group)	3	1356.712	12.1162990	0.0069952
cvd	1	1348.610	0.0147793	0.9032394
chf	1	1370.211	21.6160302	0.0000033
factor(miord)	1	1350.350	1.7542536	0.1853429
los	1	1348.736	0.1404566	0.7078275

We prune these values and look at the results:

Table 3: table detailing remaining predictors after pruning

	Df	AIC	LRT	Pr(>Chi)
	NA	1346.747	NA	NA
age	1	1403.542	58.794823	0.0000000
factor(gender)	1	1349.206	4.458412	0.0347298
hr	1	1352.978	8.230396	0.0041195
factor(bmi_group)	3	1353.218	12.470658	0.0059331
chf	1	1366.037	21.289293	0.0000039

As we can see, all predictors are now very significant. Looking at the coefficients of the model - the coefficients representing the log hazard ratios - we can see that (after exponentiating) for a unit increase in age we have a 5.96% increase in hazard. For gender, the hazard decreases by 32.53% going from male to female. For initial heart rate we see the hazard increases by 1.06% for unit increase. And for congestive heart complications, the hazard increases by 136.67% going from no to yes (which is pretty much to be expected).

```r
library(survival)


whas = read.csv("~/whas.csv", header = T)


attach(whas)
colnames(whas)


km1 = survfit(Surv(lenfol, fstat) ~ 1, conf.type = 'plain', data = whas)
plot(km1)


km2 = survfit(Surv(lenfol, fstat) ~ factor(gender), conf.type = 'plain', data = whas)
plot(km2, col = c('red', 'blue'))
survdiff(Surv(lenfol, fstat) ~ factor(gender), data = whas)


km3 = survfit(Surv(lenfol, fstat) ~ factor(bmi_group), conf.type = 'plain', data = whas)
plot(km3, col = c('red', 'blue', 'green', 'purple'))
survdiff(Surv(lenfol, fstat) ~ factor(bmi_group), data = whas)


km4 = survfit(Surv(lenfol, fstat) ~ factor(miord), conf.type = 'plain', data = whas)
plot(km4, col = c('red', 'blue'))
survdiff(Surv(lenfol, fstat) ~ factor(miord), data = whas)



#################################################################################################

fit1 = coxph(Surv(lenfol, fstat) ~ factor(miord), data = whas)
summary(fit1)

fit2 = coxph(Surv(lenfol, fstat) ~ factor(gender), data = whas)
summary(fit2)

fit3 = coxph(Surv(lenfol, fstat) ~ factor(gender) + age, data = whas)
summary(fit3)


fit4 = coxph(Surv(lenfol, fstat) ~ age, data = whas)
summary(fit4)



fita = coxph(Surv(lenfol, fstat) ~ age + factor(gender) + hr + factor(bmi_group) + cvd + chf + factor(miord)
+ los, data = whas)
drop1(fita, test = 'Chisq')

fitb = update(fita, ~. - cvd)
drop1(fitb, test = 'Chisq')

fitc = update(fitb, ~. - los)
drop1(fitc, test = 'Chisq')

fitd = update(fitc, ~. - factor(miord))
drop1(fitd, test = 'Chisq')



fitf = coxph(Surv(lenfol, fstat) ~ age + factor(gender) + hr + factor(bmi_group) + chf, data = whas)
summary(fitf)



#################################################################################################

detach(whas)
```

$\boxed{Q1}$

$$f(t) = \lambda \gamma t^{\gamma-1} e^{-\lambda t^{\gamma}} \qquad t \geq 0, \lambda > 0, \gamma > 0$$

(a)

$$S(t) = 1 - \int_0^t f(u)\, du$$

$$\left[\begin{aligned} u &= e^{-\lambda t^{\gamma}} \\ du &= -\lambda \gamma t^{\gamma-1} e^{-\lambda t^{\gamma}}\, dt \end{aligned}\right]$$

$$= 1 + \left[e^{-\lambda u^{\gamma}}\right]_0^t$$

$$= 1 + \left[e^{-\lambda t^{\gamma}} - 1\right]$$

$$= \underline{e^{-\lambda t^{\gamma}}}$$

(b)

$$h(t) = \frac{f(t)}{S(t)} = \underline{\lambda \gamma t^{\gamma-1}}$$

(c)

$$S(t) \leq 0.5$$

$$e^{-\lambda t^{\gamma}} \leq 0.5$$

$$-\lambda t^{\gamma} \leq -\ln 2$$

$$t^{\gamma} \geq \ln 2 / \lambda$$

$$\therefore \quad t \geq \left(\frac{\ln 2}{\lambda}\right)^{1/\gamma} = \underline{\left\lfloor \left(\frac{\ln 2}{\lambda}\right)^{\frac{1}{\gamma}} \right\rfloor}$$

(d) $\int_0^\infty t\, f(t)\, dt = \int_0^\infty t \lambda \gamma t^{\gamma-1} e^{-\lambda t^\gamma} dt$

$$\left[\begin{aligned} \lambda t^\gamma &= u \\ t^\gamma &= u/\lambda \\ t &= \left(u/\lambda\right)^{1/\gamma} \\ \lambda \gamma t^{\gamma-1} dt &= du \end{aligned}\right]$$

$\therefore \int_0^\infty t \lambda \gamma t^{\gamma-1} e^{-\lambda t^\gamma} dt = \int_0^\infty \left(\frac{u}{\lambda}\right)^{\frac{1}{\gamma}} e^{-u} du$

$= \int_0^\infty \left(\frac{1}{\lambda}\right)^{\frac{1}{\gamma}} . u^{\frac{1}{\gamma}} e^{-u} du$

$= \lambda^{-1/\gamma} \int_0^\infty u^{\frac{1}{\gamma}} e^{-u} du$

$= \lambda^{-1/\gamma} \Gamma\left(\frac{1}{\gamma} + 1\right)$

$$= \frac{\Gamma\left(\frac{1}{\gamma}+1\right)}{\lambda^{1/\gamma}}$$

Q2

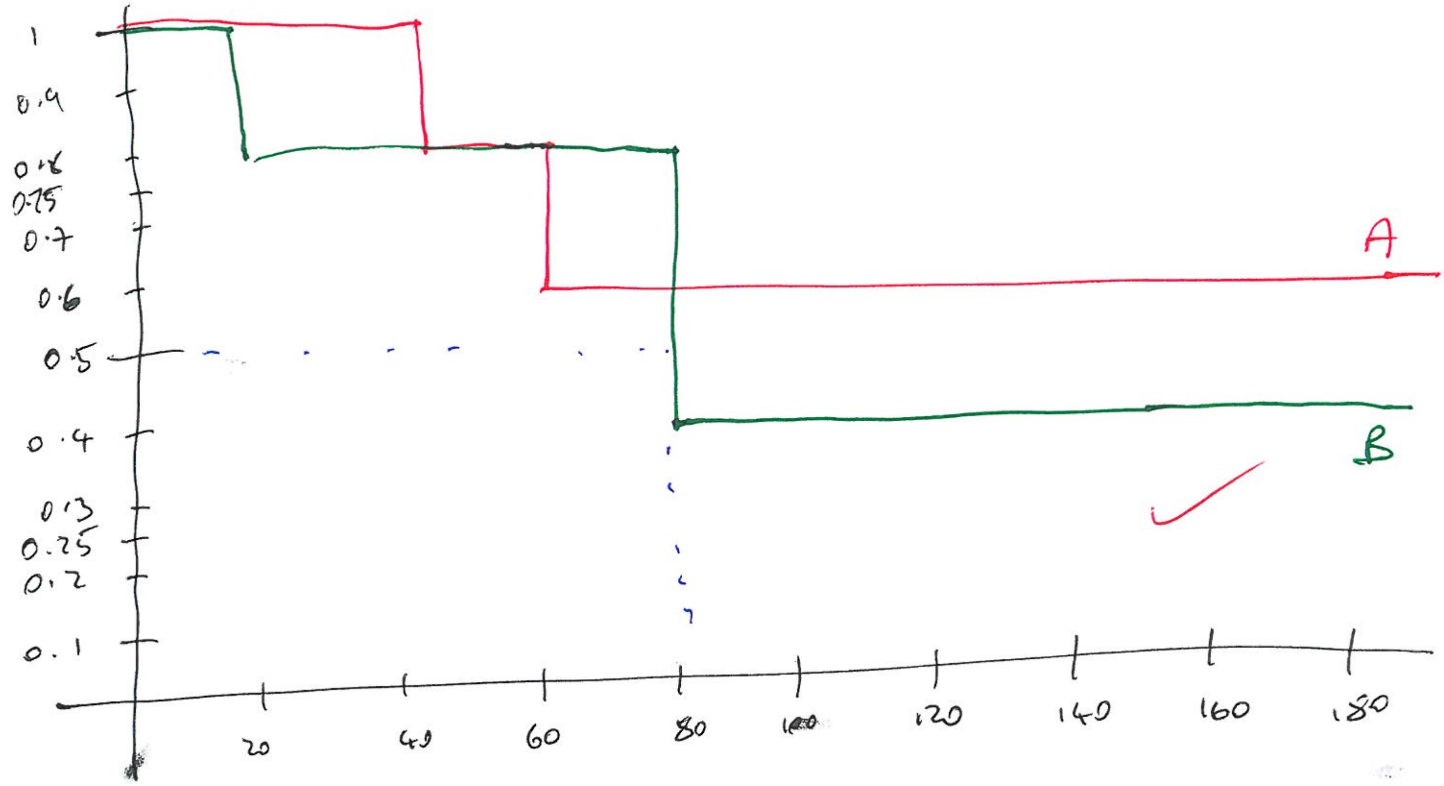
a

A

t_j	n_j	d_j	c_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
					1
0	7	0	0	1	1
5	7	0	1	1	1
24	6	0	1	1	4/5
44	5	1	0	4/5	3/5
58	4	1	0	3/4	–
59	3	0	1	–	–
150	2	0	1	–	–
161	1	0	1	–	–

B

t_j	n_j	d_j	c_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
					1
0	7	0	0	1	1
3	6	0	1	1	4/5
9	5	1	0	4/5	–
26	4	0	1	–	–
49	3	0	1	–	2/5
73	2	1	0	1/2	–
65	1	0	1	–	



median for A :

median for B : 73

t_i	d_{1j}	d_{2j}	n_{1j}	n_{2j}	n_j	d_j	e_{1j}	u_{1j}	v_{1j}
3	0	0	7	6	13	0	0	0	0
5	0	0	6	6	12	0	0	0	0
9	0	1	6	5	11	1	6/11	-6/11	0.248
24	0	0	6	5	11	0	0	0	0
26	0	0	6	4	10	0	0	0	0
44	1	0	5	4	9	1	5/9	4/9	0.247
49	0	0	5	3	8	0	0	0	0
58	[illegible]	0	4	3	7	1	4/7	3/7	0.245
59	0	0	3	3	6	0	0	0	0
73	0	1	3	2	5	1	3/5	-3/5	0.24
85	0	0	3	1	4	0	0	0	0
150	0	0	2	1	3	0	0	0	0
161	0	0	1	1	2	0	0	0	0
								-0.2724	0.98

$$\frac{u_e^2}{Var(u_e)} = 0.075$$

H_0 : survivor functions are the same

H_a : survivor functions are not the same

$$\frac{u_e^2}{Var(u_e)} \sim \chi^2_1 \qquad \text{p-value} = 0.784$$

$\therefore$ fail to reject the null hypothesis