Homework 8

Piotr Mankowski

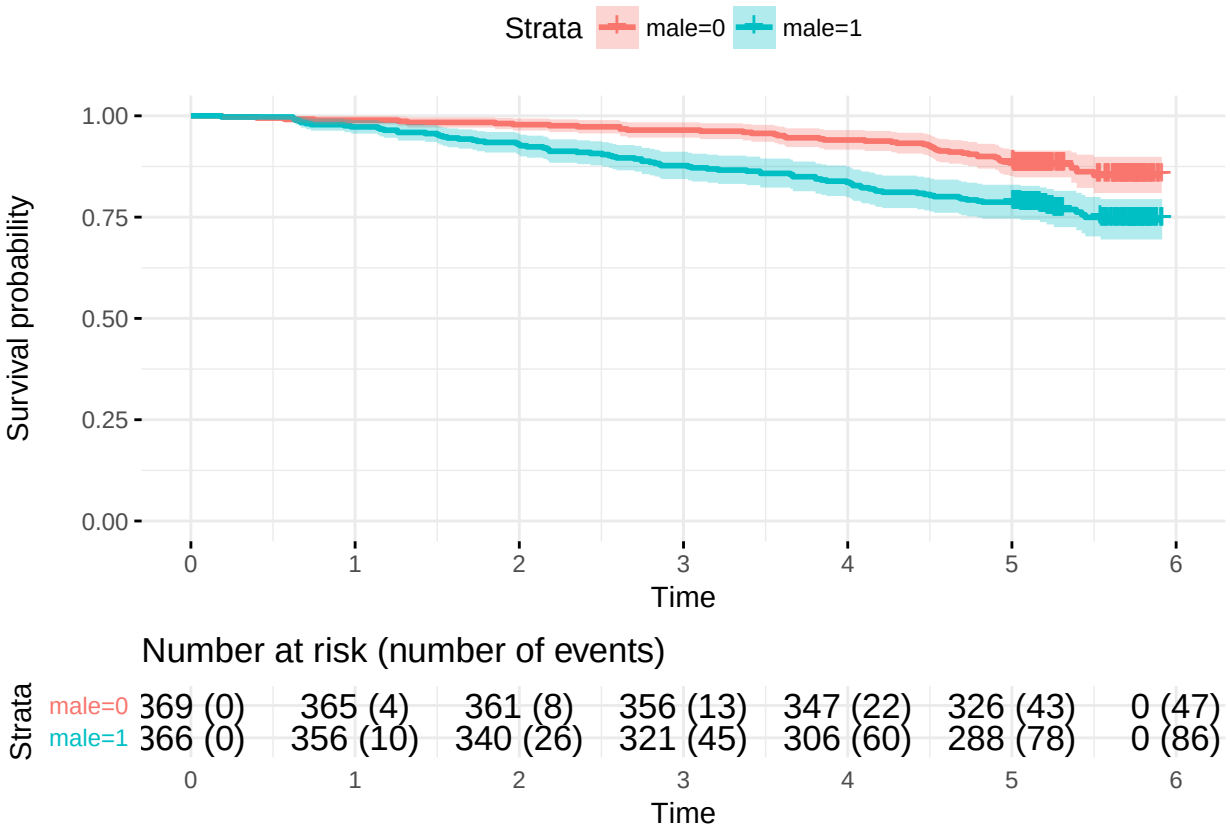
March 9, 2018

1.

Suppose we are interested in any association between risk of all-cause mortality and sex using the Kaplan-Meier estimator of the survival function. Estimate survival functions for the two sex groups using the Kaplan-Meier estimator.

a.

Provide a plot with the Kaplan-Meier estimated survival functions for the two sex groups. The two Kaplan-Meier curves should appear on the same plot. Also briefly comment on any differences/similarity of the survival curves.



In this question, we're looking at the survival of individuals in the MRI study, and comparing the survival curves of males and females. The participants were followed from the start of the study until either their death, or until September 16th, 1997; all participants whose death was not encountered had at least 5 years of follow-up, resulting in a right-censored dataset. The curves are similar, with a clustering of censoring events between years 5 and 6. This is expected, since recruitment likely happened over a ~1 year period, and all follow-ups were finished on the one date. The Kaplan-Meier Plot suggests that females in this population

had overall higher survival rates than males, which also makes sense logically, as women tend to have longer life-spans.

b.

Is there an association between risk of all-cause mortality and sex based on the Kaplan-Meier survival estimates? Explain and provide appropriate statistical evidence supporting your reasoning.

From looking at the graph, there does seem to be evidence of an association between mortality and sex; The two KM curves diverge, and have non-overlapping confidence itervals. Using the Mantel-Haenszel Log-rank test to determine if the difference between the survival curves is statistically significant, we find the p-value testing the null hypothesis of equality between the curves vs. the alternate hypothesis that the curves are not equal to be 0.000156. This result suggests that, at $\alpha = 0.05$, we should reject the null hypothesis, supporting our observation of the KM curves that there is a statistically-significant association between mortality and sex in this population.

2.

Now suppose we are interested in any association between risk of all-cause mortality and sex using a Cox proportional hazards regression model. Perform a Cox proportional hazards regression analysis of risk of mortality with sex as a predictor.

a.

Provide an interpretation of the exponentiated slope for sex in your proportional hazards regression model.

The exponentiated slope of our regression model has a value of 1.96, and represents the hazard ratio for death between males and females, with males having a 96% higher instantenous death risk.

b.

Provide full inference for an association between risk of mortality and sex from the Cox proportional hazards regression model in part a.

Methods: We investigated the association between the risk of all-cause mortality and sex for the participants of the MRI study by fitting a Cox proportional hazards regression model to survival data from the MRI study. We report the exponentiated slope for sex to estimate the hazard ratio between the two sex groups, and compute a 95% confidence interval for this estimate. We also report the the p-value for the null hypothesis that the exponentiated slope is equal to 1 (signifying equal hazard for the two groups) vs. an alternate that the slope is not equal to 1 and males and females have statistically different mortality hazards.

Results: From proportional hazards regression analysis, we estimate that the risk of death for males is 96.2% higher than for females in our study population. This estimate is highly statistically significant with a p-value of 0.000206. A 95% CI of $(1.374, 2.8)$ suggests that our results would not be unusual if the true risk of death for males was betwen 37.4% and 180% higher than for females.

c.

Compare the risk of mortality and sex association results from the Cox proportional hazards regression model to the association results in problem 1 obtained using the Kaplan-Meier method. Briefly discuss any differences in assumptions between the two methods.

The Kaplan-Meier method for survival analysis and the Log-rank test used in question 1 are non-parametric approaches that do not aim to fit any parameters, while the Cox proportional hazards regression analysis from this question is semi-parametric: the model fits a slope parameter to represent the hazard ratio, but does not fit a paramter for the baseline hazard.

Both methods agree that there's a statistically significant association between sex and mortality in the study population, and give similar p-values for this result: 0.000156 for the Log-rank test of KM survival estimates, and 0.000206 for Cox proportional hazards regression.

The Cox proportional hazards model provides an interpretable fitted estimate for the slope parameter that estimates what the actual difference in the risk of mortality is between the two groups. Being non-parametric, the KM approach does not provide such information.

3.

Now conduct a Cox proportional hazards regression analysis for risk of all-cause mortality with both sex and age at the time of study enrollment included as predictors.

a.

Provide an interpretation of the exponentiated slope for sex in your Cox proportional hazards regression model.

The exponentiated sex slope of our regression model has a value of 1.903, and represents the hazard ratio for death between males and females of the same age, with males having a 90.3% higher instantenous death risk.

b.

Provide an interpretation of the exponentiated slope for age in your Cox proportional hazards regression model.

The exponentiated age slope of our regression model has a value of 1.07, and represents the hazard ratio for death between two groups of the same sex, but differing by one year in age. The older individuals are estimated to have a 7% higher risk of death than the younger individuals.

c.

Provide full inference for an association between risk of all-cause mortality and sex with the Cox proportional hazards regression model.

Methods: We investigated the association between the risk of all-cause mortality and sex for same-age participants of the MRI study by fitting a Cox proportional hazards regression model using Huber-white estimates of the standard error to the survival data. We report the exponentiated slope for sex to estimate the hazard ratio between individuals with the same age but different sexes, and compute a 95% confidence interval for this estimate. We also report the the p-value for the null hypothesis that the exponentiated slope is equal to 1 (signifying equal hazard for the two groups) vs. an alternate that the slope is not equal to 1, and males and females of the same age have a statistically different mortality risk.

Results: From proportional hazards regression analysis, we estimate that the risk of death for males is 90.3% higher than for females of the same age in our study population. This estimate is highly statistically significant with a p-value of 0.000403. A 95% CI of $(1.33, 2.72)$ suggests that our results would not be unusual if the true risk of death for males was betwen 33% and 172% higher than for females.

d.

Provide full inference for an association between risk of all-cause mortality and age with the Cox proportional hazards regression model.

Methods: We investigated the association between the risk of all-cause mortality and age for same-sex participants of the MRI study by fitting a Cox proportional hazards regression model using Huber-white estimates of the standard error to the survival data. We report the exponentiated slope for age to estimate the hazard ratio between individuals of the same sex but 1 year apart in age, and compute a 95% confidence interval for this estimate. We also report the the p-value for the null hypothesis that the exponentiated age slope is equal to 1 (signifying equal hazard for the two groups) vs. an alternate that the slope is not equal to 1, and age is associated with mortality.

Results: From proportional hazards regression analysis, we estimate that, for two groups of the same sex but differing by 1 year in age, the risk of death is 7% higher for the older group. This estimate is highly statistically significant with a p-value of 1.07×10^{-6}. A 95% CI of $(1.040, 1.095)$ suggests that our results would not be unusual if the true risk of death for same-sex individuals but 1 year of age apart was between 4% and 9.5% higher for the older group.

e.

Does age at the time of study enrollment confound the association between risk of all-cause mortality and sex? Explain and provide evidence to support your reasoning.

In this analysis, our outcome variable is all-cause mortality risk, and our predictor of interest is sex. Although we show the age covariate to be associated with all-cause mortality and likely causal in the real world through a different pathway than sex, there is no evidence for a significant association between age and our predictor of interest, sex. A simple linear regression testing this association has a p-value of 0.426, suggesting the association is not statistically significant. This result, in turn, suggests that age does not confound the association in question.

4.

Now perform a Cox proportional hazards regression analysis of all-cause mortality with creatinine, age, sex, and indicator of ever smoked included as predictors.

a.

Provide full inference for an association between risk of all-cause mortality and creatinine with the proportional hazards regression model.

Methods: We investigated the association between the risk of all-cause mortality and serum creatinine levels for individuals of the same sex, age, and smoking status by fitting a Cox proportional hazards regression model using Huber-white estimates of the standard error to the MRI data. We report the exponentiated slope for creatinine to estimate the hazard ratio between individuals of the same sex, age, and smoking status, but with values of creatinine 1 mg/dl apart, and compute a 95% confidence interval for this estimate. We also report the the p-value for the null hypothesis that the exponentiated creatinine slope is equal to 1 (signifying no association between mortality risk and creatinine) vs. an alternate that the slope is not equal to 1.

Results: From proportional hazards regression analysis, we estimate that, for two groups of the same sex, age, and smoking status but 1 mg/dl apart in serum creatinine levels, the risk of death is 249% higher for the higher-creatinine group. This estimate is highly statistically significant with a p-value of 1.04×10^{-8}. A 95% CI of $(2.27, 5.35)$ suggests that our results would not be unusual if the true risk of death was between 127% and 434% higher for the higher-creatinine group.

b.

Provide full inference for an association between risk of all-cause mortality and smoking with the proportional hazards regression model. 1.335 0.9202 1.936 0.128

Methods: We investigated the association between the risk of all-cause mortality and smoking status for individals of the same sex, age, and serum creatinine levels by fitting a Cox proportional hazards regression model using Huber-white estimates of the standard error to the MRI data. We report the exponentiated slope for age to estimate the hazard ratio between individuals of the same sex, creatinine level, and smoking status, but 1 year apart in age, and compute a 95% confidence interval for this estimate. We also report the the p-value for the null hypothesis that the exponentiated age slope is equal to 1 (signifying no association between mortality risk and age) vs. an alternate that the slope is not equal to 1.

Results: From proportional hazards regression analysis, we estimate that, for two groups of the same sex, serum creatinine level, and smoking status but 1 year apart in age, the risk of death is 33.5% higher for the older group. This estimate is not statistically significant at $\alpha = 0.05$, with a p-value of 0.128. A 95% CI of $(0.920, 1.936)$ suggests that our results would not be unusual if the true risk of mortality was between 8% lower and 93.6% higher for the older group.

5.

Now perform a Cox proportional hazards regression analysis and provide inference on whether sex modifies an age and ever smoked adjusted association between mortality and creatinine. Explain and provide full inference supporting your reasoning.

Methods: We investigated the association between the risk of all-cause mortality and blood creatinine levels adjusted for age and smoking status, and determined if there's evidence this association is modified by sex. We performed a Cox proportional hazards regression, using Huber-White estimates of the standard error, on all-cause mortality survival data. In our model, we included serum creatinine as the predictor of interest, added age, smoking status, and sex as co-variates, and included a creatinine-sex interaction term to look at whether sex modifies the creatinine - mortality association. We report the slope for the creatinine-sex interaction term, and compute a 95% confidence interval for this estimate. We also report the the p-value for the null hypothesis that this slope slope is equal to 0, which signifies no interaction btween the two variables.

Results: From proportional hazards regression analysis, we estimate that the slope of the interaction term between creatinine and sex is 0.434. This estimate is not statistically significant at $\alpha = 0.05$, with a p-value of 0.340. A 95% CI of $(-0.021, 0.889)$ suggests that our results would not be unusual if the true creatinine-sex interaction term slope was between these two values.

The interaction term - which can be interpreted as the difference in slopes for creatinine between populations of males and females, with males having the larger slope - is estimated to not be 0. This would suggest sex modifies the creatinine-mortality association. However, this result is not statistically significant; we cannot reject the null hypothesis that the interaction slope is, in fact, 0 based on our regression results. Therefore, we cannot say that there's statistical evidence for effect modification.

Appendix

Setup code

```
knitr::opts_chunk$set(echo = FALSE, results = 'hide', warning = FALSE)

require(uwIntroStats)
library(ggplot2)
library(data.table)
library(survival)
library(survminer)

mri <- as.data.table(read.table("../data/mri.txt", header = TRUE))
mri[, obstime_yr:=obstime/365]
mri[, dead_at_5yr:=ifelse((obstime_yr <= 5.0 & death == 1), 1, 0), by='ptid']
mri[, white:=as.integer(race == 1)]
mri[, smoked:=as.integer(packyrs > 0)]
```

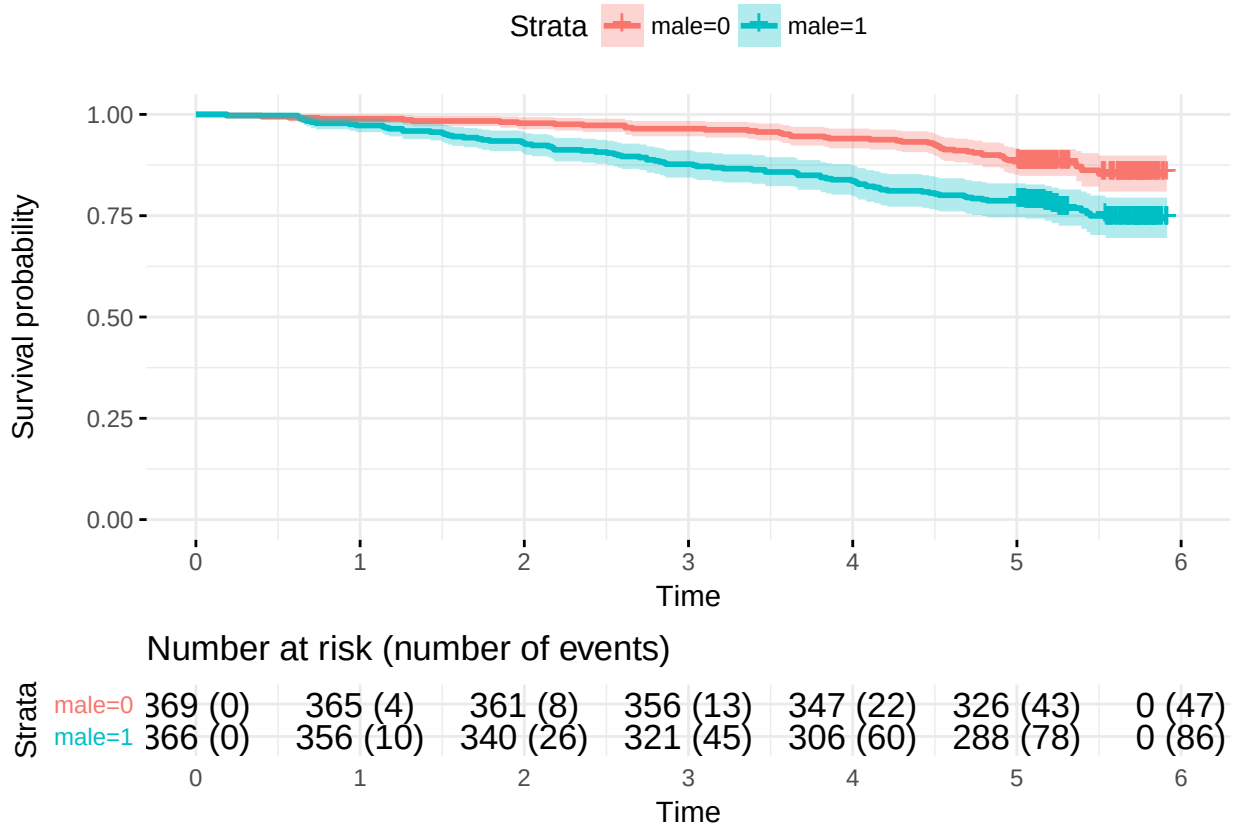
Question 1 code

```
surv <- mri[,Surv(time=obstime_yr, event=death)]
mri.df <- as.data.frame(mri)
mri.df$surv <- surv

kms <- survfit(surv ~ male, data=mri)
mri[death == 0, min(obstime_yr)]
```

```
## [1] 5.005479
```

```
ggsurvplot(kms, data=mri, risk.table = 'nrisk_cumevents',  conf.int=TRUE, ggtheme = theme_minimal())
```




```
survdiff(surv ~ male, data=mri.df)
```

```
## Call:
## survdiff(formula = surv ~ male, data = mri.df)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## male=0 369       47     68.8      6.89      14.3
## male=1 366       86     64.2      7.38      14.3
##
##  Chisq= 14.3  on 1 degrees of freedom, p= 0.000156
```

Question 2 code

```
m2 <- coxph(surv ~ male, data=mri.df)
summary(m2)
```

```
## Call:
## coxph(formula = surv ~ male, data = mri.df)
##
##   n= 735, number of events= 133
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## male 0.6739    1.9618   0.1816 3.711 0.000206 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##      exp(coef) exp(-coef) lower .95 upper .95
## male     1.962     0.5097     1.374       2.8
##
## Concordance= 0.586  (se = 0.022 )
## Rsquare= 0.019   (max possible= 0.902 )
## Likelihood ratio test= 14.45  on 1 df,   p=0.0001437
## Wald test            = 13.77  on 1 df,   p=0.0002062
## Score (logrank) test = 14.3  on 1 df,   p=0.0001559
```

Question 3 code

```
m3 <- coxph(surv ~ male + age, data=mri.df)
summary(m3)
```

```
## Call:
## coxph(formula = surv ~ male + age, data = mri.df)
##
##   n= 735, number of events= 133
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## male 0.64344   1.90301  0.18185 3.538 0.000403 ***
## age  0.06486   1.06701  0.01330 4.878 1.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## male     1.903     0.5255     1.332     2.718
## age      1.067     0.9372     1.040     1.095
##
## Concordance= 0.647  (se = 0.026 )
## Rsquare= 0.047   (max possible= 0.902 )
## Likelihood ratio test= 35.45  on 2 df,   p=2.007e-08
## Wald test            = 38.49  on 2 df,   p=4.389e-09
## Score (logrank) test = 38.86  on 2 df,   p=3.646e-09
```

```
regress("mean", age ~ male, data=mri)
```

```
##
## Call:
## regress(fnctl = "mean", formula = age ~ male, data = mri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4065 -3.7268 -0.7268  3.2732 24.2732
##
## Coefficients:
##                   Estimate  Naive SE  Robust SE    95%L      95%H
## [1] Intercept      74.41     0.2839    0.2737       73.87     74.94
## [2] male           0.3203    0.4023    0.4024      -0.4697    1.110
##                     F stat    df Pr(>F)
## [1] Intercept        73899.43 1  < 0.00005
## [2] male                 0.63 1    0.4263
##
```

```
## Residual standard error: 5.453 on 733 degrees of freedom
## Multiple R-squared:  0.0008641,  Adjusted R-squared:  -0.000499
## F-statistic: 0.6335 on 1 and 733 DF,  p-value: 0.4263
```

Question 4 code

```
m4 <- coxph(surv ~ crt + age + male + smoked, data=mri.df)
summary(m4)
```

```
## Call:
## coxph(formula = surv ~ crt + age + male + smoked, data = mri.df)
##
##   n= 732, number of events= 132
##    (3 observations deleted due to missingness)
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## crt    1.24886   3.48637  0.21817 5.724 1.04e-08 ***
## age    0.06784   1.07020  0.01393 4.871 1.11e-06 ***
## male   0.27719   1.31942  0.18966 1.462    0.144
## smoked 0.28884   1.33487  0.18978 1.522    0.128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## crt        3.486     0.2868    2.2733     5.347
## age        1.070     0.9344    1.0414     1.100
## male       1.319     0.7579    0.9098     1.913
## smoked     1.335     0.7491    0.9202     1.936
##
## Concordance= 0.669  (se = 0.026 )
## Rsquare= 0.079   (max possible= 0.901 )
## Likelihood ratio test= 60.52  on 4 df,   p=2.252e-12
## Wald test            = 71.48  on 4 df,   p=1.11e-14
## Score (logrank) test = 72.27  on 4 df,   p=7.55e-15
```

Question 5 code

```
m5_males <- coxph(surv ~ crt + age + smoked, data=mri.df[which(mri.df$male == 1), ])
m5_females <- coxph(surv ~ crt + age + smoked, data=mri.df[which(mri.df$male == 0), ])


mri.df$maleANDcrt <- mri.df$male * mri.df$crt
m5 <- coxph(surv ~ crt + age + smoked + male + maleANDcrt, data=mri.df)
m5 <- coxph(surv ~ crt * male + age + smoked, data=mri.df)

summary(m5_males)
```

```
## Call:
## coxph(formula = surv ~ crt + age + smoked, data = mri.df[which(mri.df$male ==
##     1), ])
##
##   n= 365, number of events= 85
```

```
##    (1 observation deleted due to missingness)
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## crt    1.50971   4.52544  0.29956 5.040 4.66e-07 ***
## age    0.03803   1.03876  0.01789 2.126   0.0335 *
## smoked 0.15771   1.17083  0.24105 0.654   0.5129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## crt        4.525     0.2210     2.516     8.140
## age        1.039     0.9627     1.003     1.076
## smoked     1.171     0.8541     0.730     1.878
##
## Concordance= 0.597  (se = 0.032 )
## Rsquare= 0.063   (max possible= 0.93 )
## Likelihood ratio test= 23.61  on 3 df,   p=3.015e-05
## Wald test            = 32.23  on 3 df,   p=4.687e-07
## Score (logrank) test = 30.09  on 3 df,   p=1.321e-06
```

```
summary(m5_females)
```

```
## Call:
## coxph(formula = surv ~ crt + age + smoked, data = mri.df[which(mri.df$male ==
##     0), ])
##
##   n= 367, number of events= 47
##    (2 observations deleted due to missingness)
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## crt    1.01882   2.76993  0.38005 2.681  0.00734 **
## age    0.13434   1.14379  0.02529 5.313 1.08e-07 ***
## smoked 0.54698   1.72803  0.30546 1.791  0.07335 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## crt        2.770     0.3610    1.3151     5.834
## age        1.144     0.8743    1.0885     1.202
## smoked     1.728     0.5787    0.9496     3.145
##
## Concordance= 0.72  (se = 0.043 )
## Rsquare= 0.089   (max possible= 0.771 )
## Likelihood ratio test= 34.06  on 3 df,   p=1.926e-07
## Wald test            = 36.49  on 3 df,   p=5.903e-08
## Score (logrank) test = 39.68  on 3 df,   p=1.248e-08
```

```
summary(m5)
```

```
## Call:
## coxph(formula = surv ~ crt * male + age + smoked, data = mri.df)
##
##   n= 732, number of events= 132
##    (3 observations deleted due to missingness)
##
```

```
##              coef exp(coef) se(coef)      z Pr(>|z|)
## crt       1.01098   2.74828  0.35470  2.850  0.00437 **
## male     -0.22922   0.79515  0.55582 -0.412  0.68005
## age       0.06732   1.06964  0.01401  4.805 1.55e-06 ***
## smoked    0.31027   1.36379  0.19134  1.622  0.10490
## crt:male  0.43385   1.54318  0.45506  0.953  0.34039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## crt         2.7483     0.3639    1.3713     5.508
## male        0.7952     1.2576    0.2675     2.364
## age         1.0696     0.9349    1.0407     1.099
## smoked      1.3638     0.7333    0.9373     1.984
## crt:male    1.5432     0.6480    0.6325     3.765
##
## Concordance= 0.668  (se = 0.026 )
## Rsquare= 0.081   (max possible= 0.901 )
## Likelihood ratio test= 61.46  on 5 df,   p=6.053e-12
## Wald test            = 77.14  on 5 df,   p=3.331e-15
## Score (logrank) test = 79.49  on 5 df,   p=1.11e-15
```