432 Class 24 Slides

github.com/THOMASELOVE/432-2018

2018-04-12

Preliminaries

```r
library(skimr)
library(rms)
library(survival)
library(OIsurv)
library(survminer)
library(broom)
library(tidyverse)
```

```r
survex <- read.csv("data/survex.csv") %>% tbl_df
```

Today's Agenda

- Data Visualization Examples
- Regression on Time-to-event data
    - Cox Proportional Hazards Model

Today's Visualization

- Baby Name Voyager
- Better Health Partnership's Interactive Social Determinants and Children's Health tool

Regression on Time-to-Event / Survival Outcomes

The Cox Proportional Hazards Model: An Introduction

The Cox proportional hazards (Cox regression) model fits survival data with a constant (i.e. not varying over time) covariate x to a hazard function of the form:

$$h(t|x) = h_0(t)exp[\beta_1 x]$$

where we will estimate the unknown value of β_1 and where $h_0(t)$ is the baseline hazard, which is a non-parametric and unspecified value which depends on t but not on x.

More on the Cox Model

- For particular x values, we will be able to estimate the survival function if we have an estimate of the baseline survival function, $\hat{S}_0(t)$.

The estimated survival function for an individual with covariate value x_k turns out to be

$$\hat{S}(t|x_k) = [\hat{S}_0(t)]^{exp(\beta_1 x_k)}$$

Fitting the Cox Model with `coxph`

Fitting a Cox Model in R

There are two main approaches to fitting Cox models in R.

- the `coxph` function in the `survival` package, and
- the `cph` function in the `rms` package.

We'll start with the `coxph` approach, and fit a pair of models to the `survex` data.

The `survex` data frame

The `survex.csv` file on the course website is essentially the same as a file simulated by Frank Harrell and his team[1] to introduce some of the key results from the `cph` function, which is part of the `rms` package in R.

The `survex` data includes 1,000 subjects...

- `id` = patient ID (1-1000)
- `age` = patient's age at study entry, years
- `sex` = patient's sex (Male or Female)
- `study.yrs` = patient's years of observed time in study until death or censoring
- `death` = 1 if patient died, 0 if censored.

We'll start by creating a survival object, then fitting it using `sex` as a predictor.

[1] see the rms package documentation

A Cox Model for the `survex` data using `sex`

```r
model1 <- with(survex, coxph(Surv(study.yrs, death) ~ sex))
model1
```

```
Call:
coxph(formula = Surv(study.yrs, death) ~ sex)

          coef exp(coef) se(coef)     z       p
sexMale -0.619     0.538    0.148 -4.18 2.9e-05

Likelihood ratio test=17.2  on 1 df, p=3.4e-05
n= 1000, number of events= 183
```

- This tiny summary provides an overall comparison of males to females, using a proportional hazards model.
    - The default R approach uses the "efron" method of breaking ties: other options include "breslow" and "exact".

summary(model1)

```
> summary(model1)
Call:
coxph(formula = Surv(study.yrs, death) ~ sex)

  n= 1000, number of events= 183

           coef exp(coef) se(coef)      z Pr(>|z|)
sexMale -0.6195    0.5382   0.1481 -4.184 2.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
sexMale    0.5382      1.858    0.4027    0.7194

Concordance= 0.586  (se = 0.019 )
Rsquare= 0.017   (max possible= 0.903 )
Likelihood ratio test= 17.18  on 1 df,   p=3.399e-05
Wald test            = 17.51  on 1 df,   p=2.862e-05
Score (logrank) test = 18.07  on 1 df,   p=2.129e-05
```

Interpreting the Hazard Ratio estimate

Our hazard ratio estimate is 0.54 for Males (compared to Females)

```
        exp(coef) exp(-coef) lower .95 upper .95
sexMale    0.5382      1.858    0.4027    0.7194
```

The hazard ratio is a multiplicative effect of the covariate (Male sex) on the hazard function for death.

- A hazard ratio of 1 indicates no effect
- A hazard ratio < 1 indicates a decrease in the hazard for Males as compared to Females
- A hazard ratio > 1 indicates an increase in the hazard for Males as compared to Females

Likelihood Ratio Test in more detail via anova

```
anova(model1)
```

```
Analysis of Deviance Table
 Cox model: response is Surv(study.yrs, death)
Terms added sequentially (first to last)

      loglik Chisq Df Pr(>|Chi|)
NULL -1167.8
sex  -1159.2 17.18  1  3.399e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2: age and sex

```
(model2 <- with(survex,
    coxph(Surv(study.yrs, death) ~ age + sex)
))
```

```
Call:
coxph(formula = Surv(study.yrs, death) ~ age + sex)

            coef exp(coef) se(coef)     z       p
age      0.04192   1.04281  0.00557  7.53 5.3e-14
sexMale -0.59753   0.55017  0.14821 -4.03 5.5e-05

Likelihood ratio test=69.9  on 2 df, p=6.66e-16
n= 1000, number of events= 183
```

summary(model2)

```
> summary(model2)
Call:
coxph(formula = Surv(study.yrs, death) ~ age + sex)

  n= 1000, number of events= 183

             coef exp(coef)  se(coef)      z Pr(>|z|)
age      0.041920  1.042811  0.005571  7.525 5.26e-14 ***
sexMale -0.597528  0.550170  0.148207 -4.032 5.54e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        exp(coef) exp(-coef) lower .95 upper .95
age        1.0428     0.9589    1.0315    1.0543
sexMale    0.5502     1.8176    0.4115    0.7356

Concordance= 0.688  (se = 0.023 )
Rsquare= 0.068   (max possible= 0.903 )
Likelihood ratio test= 69.93  on 2 df,   p=6.661e-16
Wald test            = 75.83  on 2 df,   p=0
Score (logrank) test = 73.33  on 2 df,   p=1.11e-16
```

Interpreting the Hazard Ratio estimate

```
        exp(coef) exp(-coef) lower .95 upper .95
age        1.0428     0.9589    1.0315    1.0543
sexMale    0.5502     1.8176    0.4115    0.7356
```

- If Harry is one year older than Steve, and both are male, then Harry's hazard of death is 1.04 times that of Steve (95% CI 1.03, 1.05). Alternatively, Steve's Hazard is 0.96 times that of Harry.
- If Harry (male) and Sally (female) are the same age, then Harry's hazard of death is 0.55 times that of Sally (95% CI 0.41, 0.74). Alternatively, Sally's hazard is 1.82 times that of Harry.

Concordance and R^2 Summaries

```
Concordance= 0.688  (se = 0.023 )
Rsquare= 0.068   (max possible= 0.903 )
```

- Concordance is only appropriate when we have at least one continuous predictor in our Cox model, in which case it assesses the probability of agreement between the survival time and the risk score generated by the (continuous) predictor or set of predictors. A value of 1 indicates perfect agreement, 0.5 is no better than chance. Our concordance = 0.69, which is a fairly typical value.
- Rsquare here is Cox and Snell's pseudo-R^2, which reflects the improvement of the model we have fit over the model with the intercept alone, as tested by the likelihood ratio test.
    - The maximum value of this statistic is often less than one, in which case R will tell you that.

Tidy the model's coefficients with `broom::tidy`

```r
tidy(model2)
```

```
     term    estimate  std.error statistic      p.value
1     age  0.04191953 0.00557055  7.525205 5.262457e-14
2 sexMale -0.59752793 0.14820706 -4.031710 5.537239e-05
     conf.low  conf.high
1  0.03100145  0.0528376
2 -0.88800842 -0.3070474
```

Glance at model summaries with `broom::glance`

```r
glance(model2)
```

```
     n nevent statistic.log  p.value.log statistic.sc
1 1000    183      69.93224 6.661338e-16     73.33081
    p.value.sc statistic.wald p.value.wald r.squared
1 1.110223e-16          75.83            0  0.067543
  r.squared.max concordance std.error.concordance    logLik
1     0.9032474   0.6877545            0.02257605 -1132.833
       AIC      BIC
1 2269.666 2276.085
```

`anova(model2)` shows sequential LR tests

```
Analysis of Deviance Table
 Cox model: response is Surv(study.yrs, death)
Terms added sequentially (first to last)

      loglik  Chisq Df Pr(>|Chi|)
NULL -1167.8
age  -1140.8 53.962  1  2.044e-13 ***
sex  -1132.8 15.970  1  6.435e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing the Key Assumption: Proportional Hazards

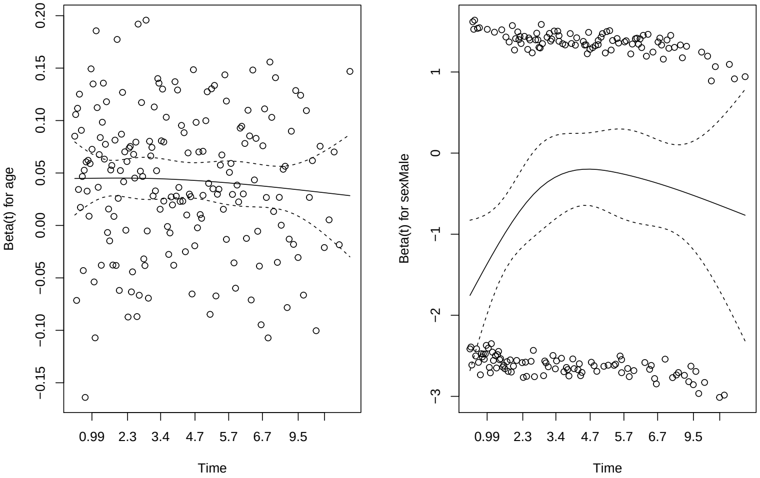
```r
cox.zph(model2, transform = "km", global = TRUE)
```

```
            rho chisq      p
age     -0.0556 0.464 0.4956
sexMale  0.1345 3.294 0.0696
GLOBAL       NA 3.880 0.1437
```

The *p* values show whether the interaction between the specified covariate and time is significant.

- A significant effect here is an indication of trouble with the PH assumption.

Plotting the `cox.zph` results



Plotting the `cox.zph` results (code)

```
par(mfrow = c(1,2))
plot(cox.zph(model2, transform = "km", global = TRUE))
par(mfrow = c(1,1))
```

- If the proportional hazards assumption is appropriate, then we should see a slope of essentially zero in each such plot.
- A slope that is seriously different from zero suggests a violation of the proportional hazards assumption.
- Here, we may have an issue with the assumption of PH in `sex`.
  - If we did, we'd either add a non-linear term (if `sex` was continuous), or use a different kind of survival model.

Building a Cox Model with `cph` from the `rms` package

Building `model2` using the `cph` function

```
d <- datadist(survex)
options(datadist="d")

S <- Surv(time = survex$study.yrs, event = survex$death)

mod2 <- cph(S ~ age + sex,
            data = survex,
            x = TRUE, y = TRUE, surv = TRUE)
```

Looking at `mod2`

```
Cox Proportional Hazards Model

 cph(formula = S ~ age + sex, data = survex, x = TRUE, y = TRU
     surv = TRUE)

                    Model Tests       Discrimination
                                         Indexes
 Obs       1000    LR chi2      69.93    R2       0.075
 Events     183    d.f.             2    Dxy      0.376
 Center 1.6933     Pr(> chi2) 0.0000    g        0.675
                   Score chi2   73.33    gr       1.965
                   Pr(> chi2) 0.0000

          Coef    S.E.   Wald Z Pr(>|Z|)
 age       0.0419 0.0056  7.53  <0.0001
 sex=Male -0.5975 0.1482 -4.03  <0.0001
```

Validation of `mod2` Summaries

```
set.seed(432109); validate(mod2)
```

```
      index.orig training   test optimism index.corrected
Dxy       0.3755   0.3830 0.3712   0.0119          0.3636
R2        0.0748   0.0788 0.0736   0.0052          0.0696
Slope     1.0000   1.0000 0.9694   0.0306          0.9694
D         0.0295   0.0313 0.0290   0.0022          0.0273
U        -0.0009  -0.0009 0.0005  -0.0014          0.0006
Q         0.0304   0.0321 0.0285   0.0036          0.0267
g         0.6753   0.6997 0.6709   0.0288          0.6465
       n
Dxy    40
R2     40
Slope  40
D      40
U      40
Q      40
```

ANOVA on mod2

```r
anova(mod2)
```

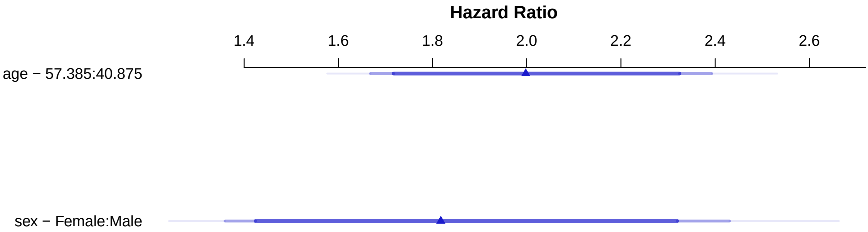
```
                Wald Statistics          Response: S

 Factor     Chi-Square d.f. P
 age        56.63      1    <.0001
 sex        16.25      1    1e-04
 TOTAL      75.83      2    <.0001
```

Effect Sizes via cph for mod2

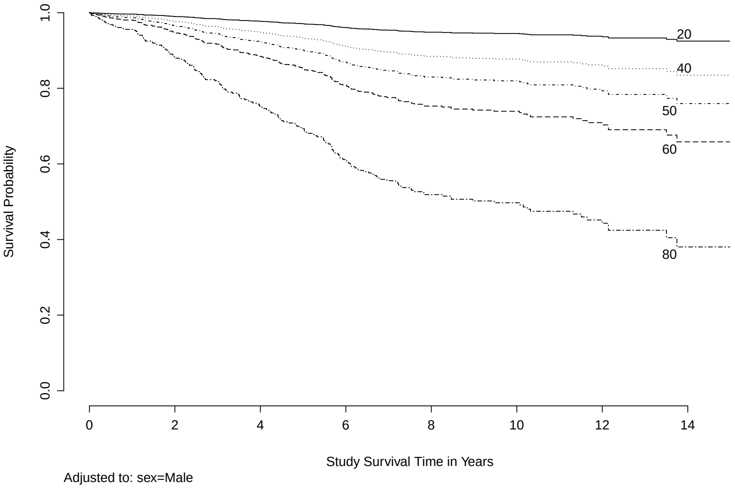
```r
summary(mod2)
```

```
             Effects              Response : S 

 Factor                Low    High   Diff. Effect  S.E.
 age                   40.875 57.385 16.51 0.69209 0.09197
  Hazard Ratio         40.875 57.385 16.51 1.99790      NA
 sex - Female:Male      2.000  1.000    NA 0.59753 0.14821
  Hazard Ratio          2.000  1.000    NA 1.81760      NA
 Lower 0.95 Upper 0.95
 0.51183    0.87235
 1.66830    2.39250
 0.30705    0.88801
 1.35940    2.43030
```

plot(summary(mod2))



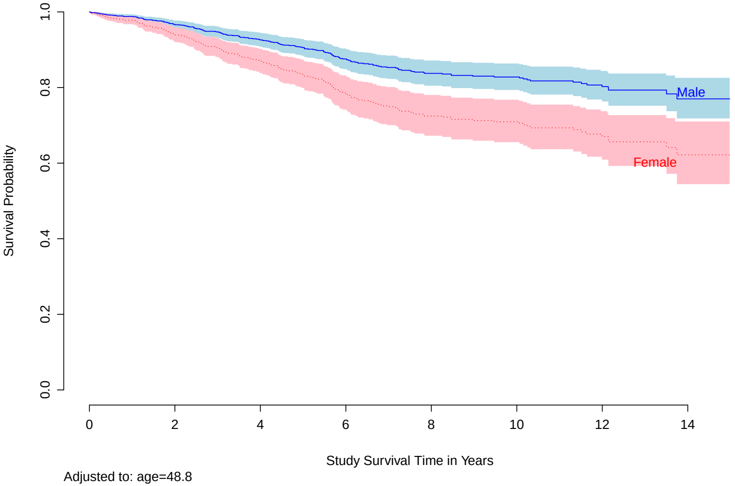
Comparing Survival for males in `mod2` at various ages



Code for prior slide

```r
survplot(mod2, age = c(20, 40, 50, 60, 80),
         time.inc=2,
         type="kaplan-meier",
         xlab="Study Survival Time in Years")
```

Comparing Survival by sex in `mod2` at median age (49)

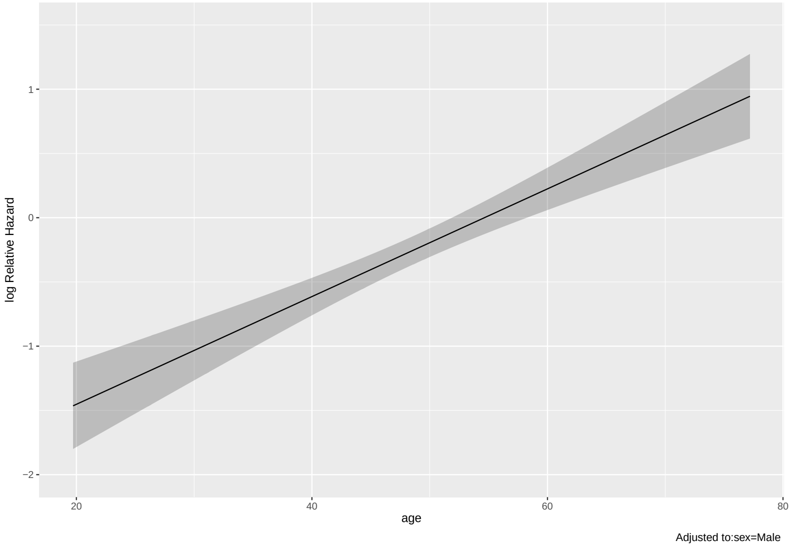


Code for prior slide

```
survplot(mod2, sex = c("Male", "Female"),
         time.inc=2,
         conf.int = TRUE,
         col = c("blue","red"),
         col.fill = c("lightblue", "pink"),
         type="kaplan-meier",
         xlab="Study Survival Time in Years")
```

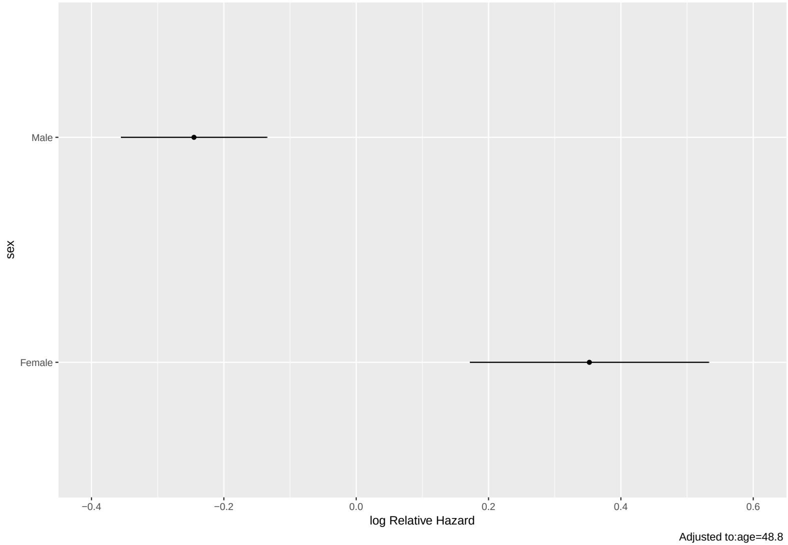
Plotting the age effect implied by mod2

```
ggplot(Predict(mod2, age))
```



Plotting the sex effect implied by mod2

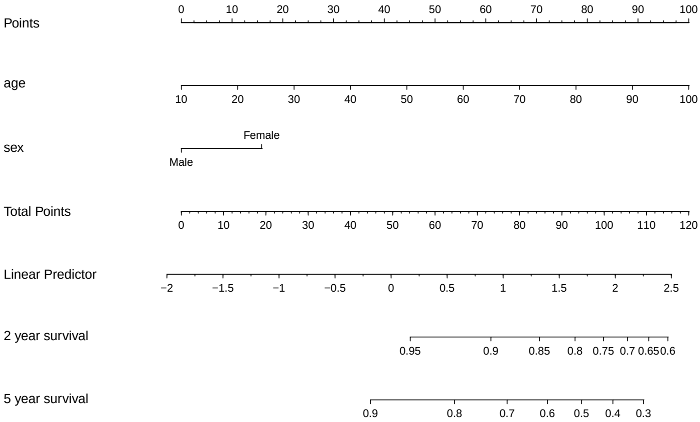
```
ggplot(Predict(mod2, sex))
```



mod2 Nomogram (code)

```
sv <- Survival(mod2)
surv2 <- function(x) sv(2, lp = x)
surv5 <- function(x) sv(5, lp = x)

plot(nomogram(mod2,
              fun = list(surv2, surv5),
              funlabel = c("2 year survival",
                  "5 year survival")))
```

Resulting `mod2` Nomogram



Model 3, with a spline in age and age-sex interaction

```
d <- datadist(survex)
options(datadist="d")

S <- Surv(time = survex$study.yrs, event = survex$death)

mod3 <- cph(S ~ rcs(age,4) + catg(sex) + age %ia% sex,
            data = survex,
            x = TRUE, y = TRUE, surv = TRUE)
```

Looking at `mod3`

```
> mod3
Cox Proportional Hazards Model

 cph(formula = S ~ rcs(age, 4) + catg(sex) + age %ia% sex, data = survex,
     x = TRUE, y = TRUE, surv = TRUE)

                      Model Tests       Discrimination
                                           Indexes
 Obs         1000    LR chi2      79.06    R2       0.084
 Events       183    d.f.             5    Dxy      0.379
 Center -0.6562      Pr(> chi2) 0.0000    g        0.797
                     Score chi2   85.72    gr       2.219
                     Pr(> chi2) 0.0000

               Coef    S.E.   Wald Z Pr(>|Z|)
 age           -0.0243 0.0297 -0.82  0.4139
 age'           0.2048 0.0774  2.65  0.0082
 age''         -0.7455 0.2706 -2.76  0.0059
 sex=Male      -1.2391 0.6816 -1.82  0.0691
 age * sex=Male  0.0108 0.0120  0.89  0.3711
```

Validate summary statistics in mod3

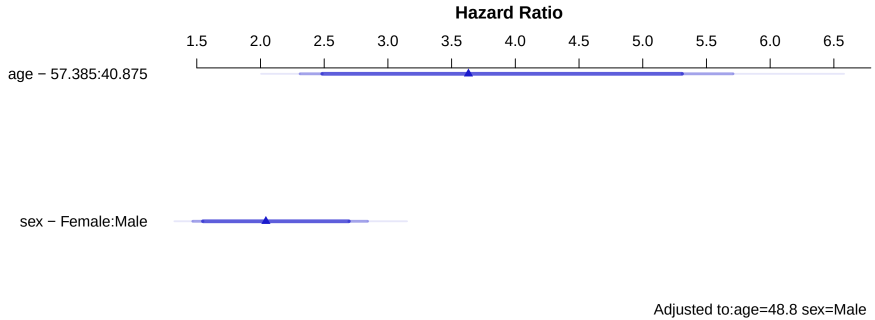
```
set.seed(432301); validate(mod3)
```

```
      index.orig training   test optimism index.corrected
Dxy       0.3790   0.3866 0.3742   0.0124          0.3667
R2        0.0842   0.0886 0.0794   0.0093          0.0749
Slope     1.0000   1.0000 0.9482   0.0518          0.9482
D         0.0334   0.0352 0.0314   0.0038          0.0296
U        -0.0009  -0.0009 0.0008  -0.0016          0.0008
Q         0.0343   0.0361 0.0306   0.0054          0.0288
g         0.7969   0.8147 0.7632   0.0515          0.7454
       n
Dxy   40
R2    40
Slope 40
D     40
U     40
Q     40
```

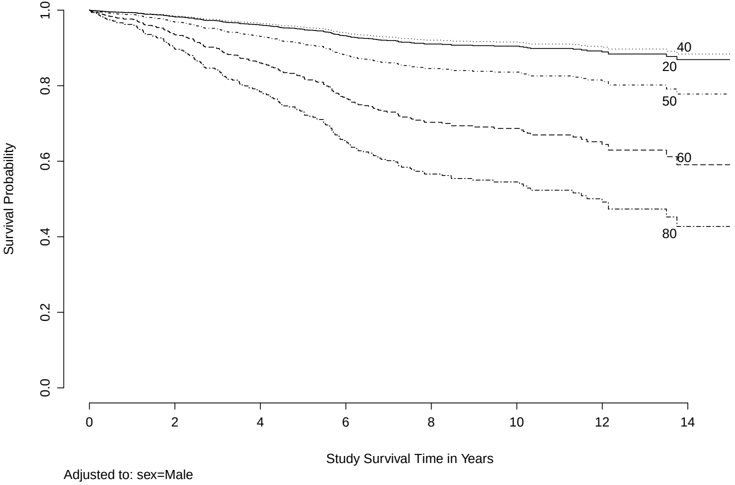
summary(mod3)

```
            Effects              Response : S

 Factor                 Low    High   Diff. Effect  S.E.
 age                    40.875 57.385 16.51 1.28980 0.2307
  Hazard Ratio          40.875 57.385 16.51 3.63210     NA
 sex - Female:Male       2.000  1.000    NA 0.71417 0.1682
  Hazard Ratio           2.000  1.000    NA 2.04250     NA
 Lower 0.95 Upper 0.95
 0.83763    1.7420
 2.31090    5.7085
 0.38450    1.0438
 1.46890    2.8401

Adjusted to: age=48.8 sex=Male
```

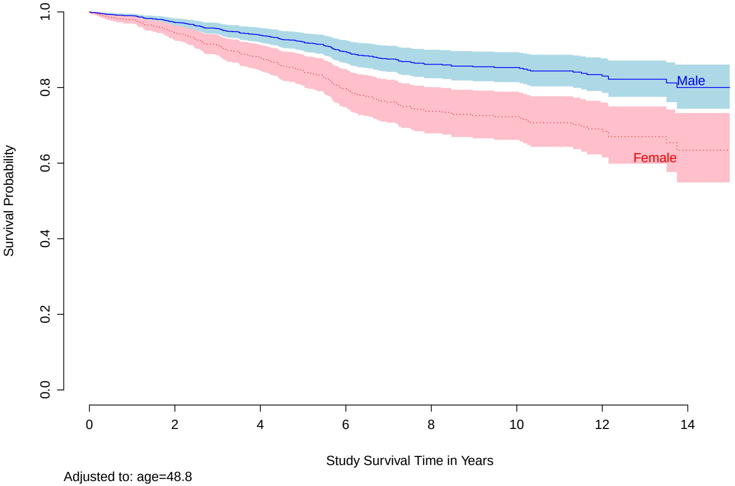
plot(summary(mod3))



Comparing Survival for males in `mod3` at various ages

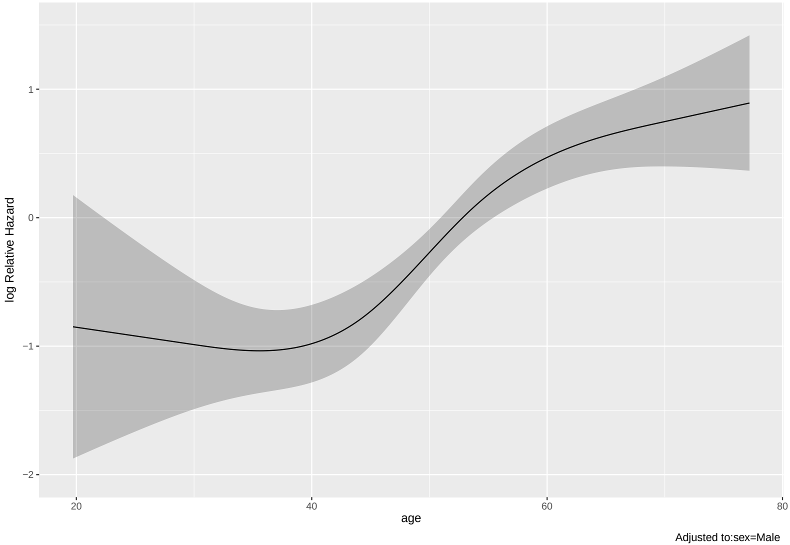


Comparing Survival by sex in `mod3` at median age (49)



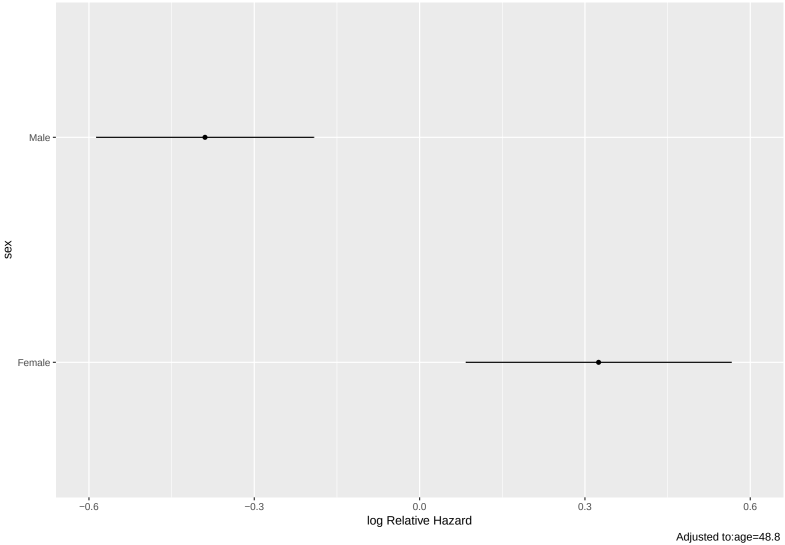
Plotting the `age` effect implied by `mod2`

```
ggplot(Predict(mod3, age))
```

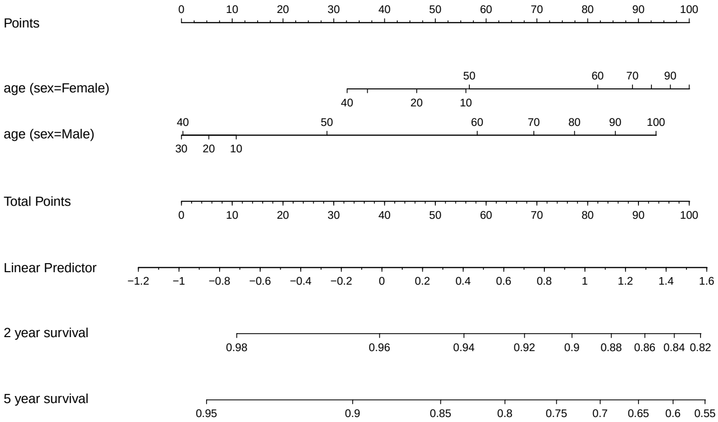


Plotting the `sex` effect implied by `mod2`

```
ggplot(Predict(mod3, sex))
```



`mod3` Nomogram

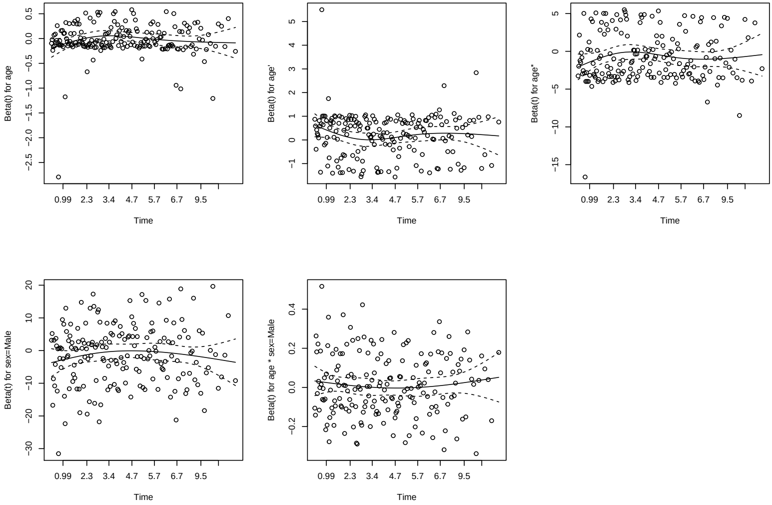


Checking the Proportional Hazards Assumption

```
cox.zph(mod3, transform = "km", global = TRUE)
```

```
                   rho   chisq     p
age            0.01848 0.04542 0.831
age'          -0.02293 0.07402 0.786
age''          0.01694 0.04123 0.839
sex=Male       0.03450 0.20930 0.647
age * sex=Male -0.00415 0.00302 0.956
GLOBAL              NA 4.16926 0.525
```

Plots for PH Assumption



Next Time

Another survival analysis example