Model Fitting and Diagnostics for the Cox Model

Steve Simon

Abstract

Lecture 4. Model fitting and diagnostics for the Cox model. In this lecture, you will work with more complex forms of the Cox model with multiple predictor variables. You'll include covariates in the Cox model to produce risk adjusted survival curves. You will also assess the linearity assumptions using Martingale residuals and splines.

Advantages of a Multivariate Model

1. Your predictions are better with two (or more) independent variables.
2. You can use covariates to make risk adjustments.
3. You can explore interactions among variables.

Review the Whas500 Data Det

```
##   id age gender hr sysbp diasbp      bmi cvd afb
## 1  1  83   Male 89   152     78 25.54051  No Yes
## 2  2  49   Male 84   120     60 24.02398  No  No
##   sho chf av3     miord     mitype year
## 1  No  No  No Recurrent Non Q-wave <NA>
## 2  No  No  No     First     Q-wave <NA>
##    admitdate    disdate      fdate los dstat
## 1 01/13/1997 01/18/1997 12/31/2002   5 Alive
## 2 01/19/1997 01/24/1997 12/31/2002   5 Alive
##   lenfol fstat time_yrs
## 1   2178 Alive 5.963039
## 2   2172 Alive 5.946612
```

Model Fitting Strategies

1. Fit univariate models first.
2. Add variables one at a time or in very small batches.
3. Look at interactions and nonlinearities last.

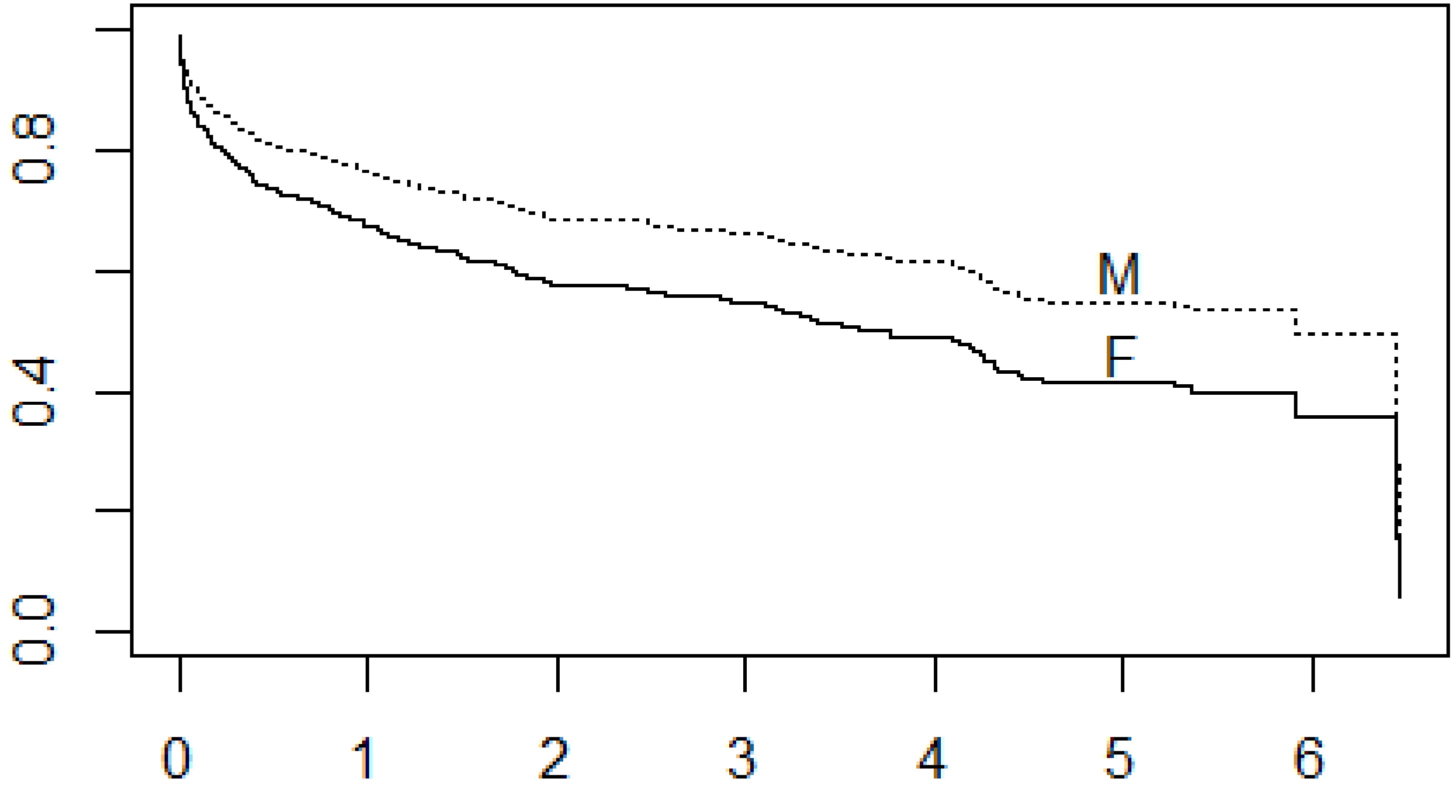
Univariate Model for Age

```
##   term   hr p.value     conf.int
## 1  age 1.07   0.001 1.06 to 1.08
```

Univariate Model for Gender

```
##          term   hr p.value      conf.int
## 1 genderFemale 1.46   0.006 1.12 to 1.92
```

Estimated Survival by Gender



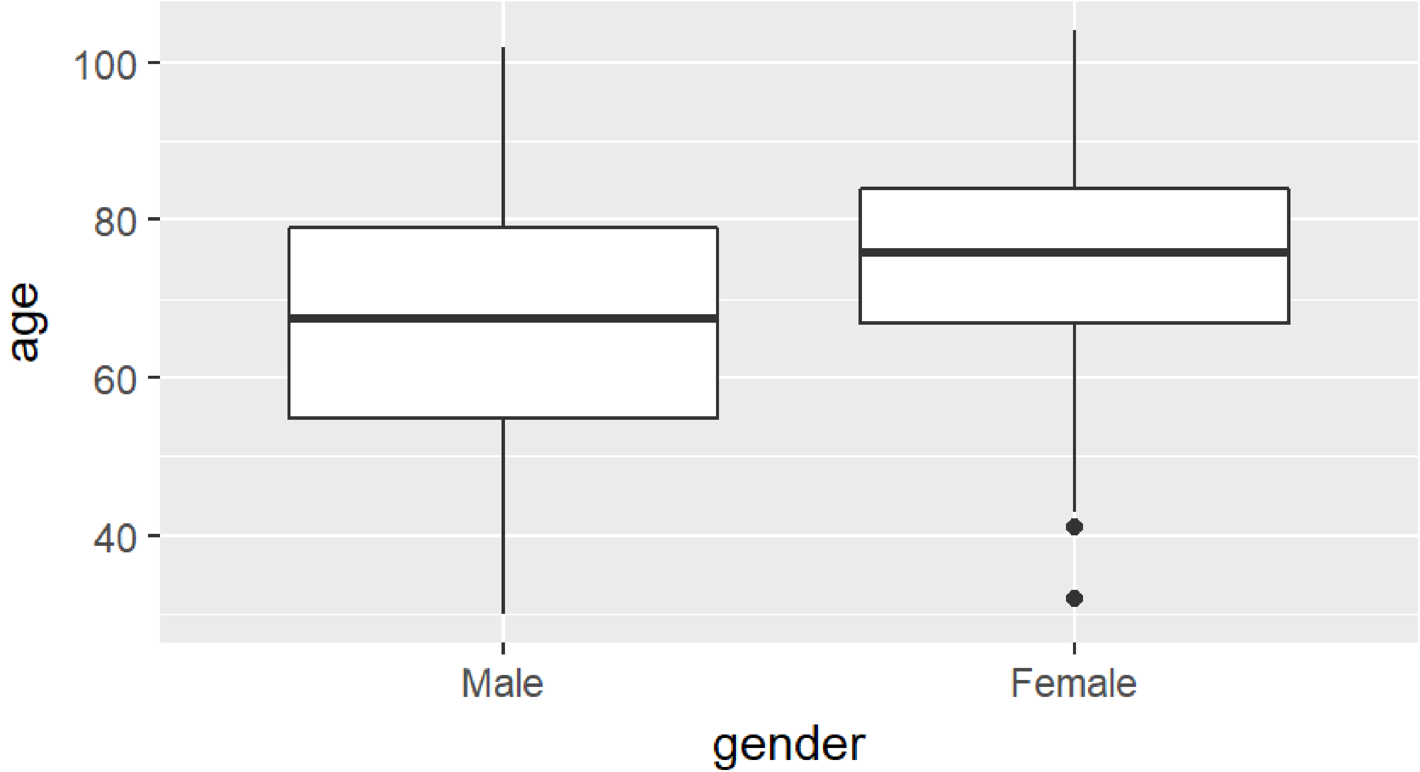
Model with Age and Gender

```
##           term   hr p.value     conf.int
## 1          age 1.07   0.637 1.06 to 1.08
## 2 genderFemale 0.94   0.637 0.71 to 1.23
```

What is Happening Here?

The average age across all subjects is 69.8, but the averages by gender are quite different. For males, the average age is 66.6, but for females, the average age is 74.7.

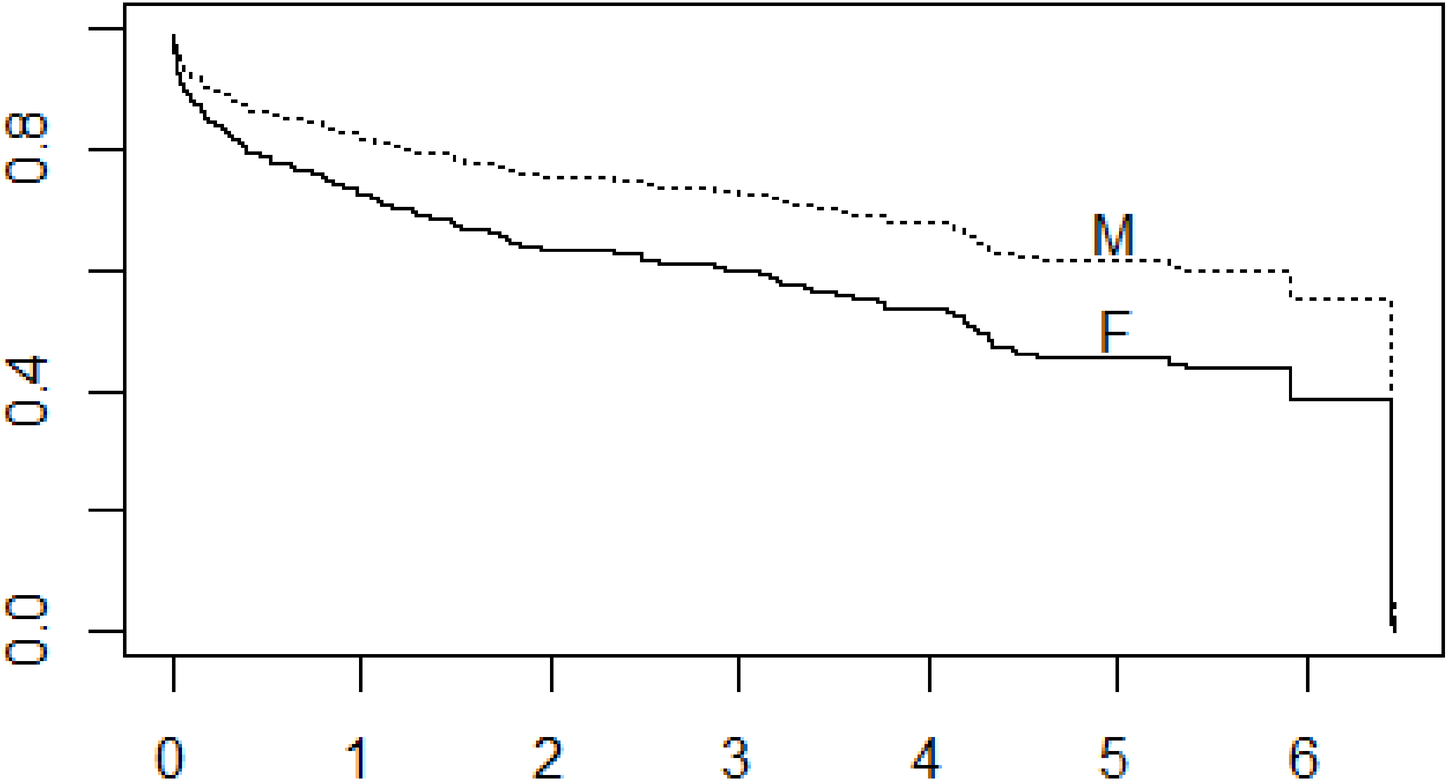
Boxplots of Age by Gender



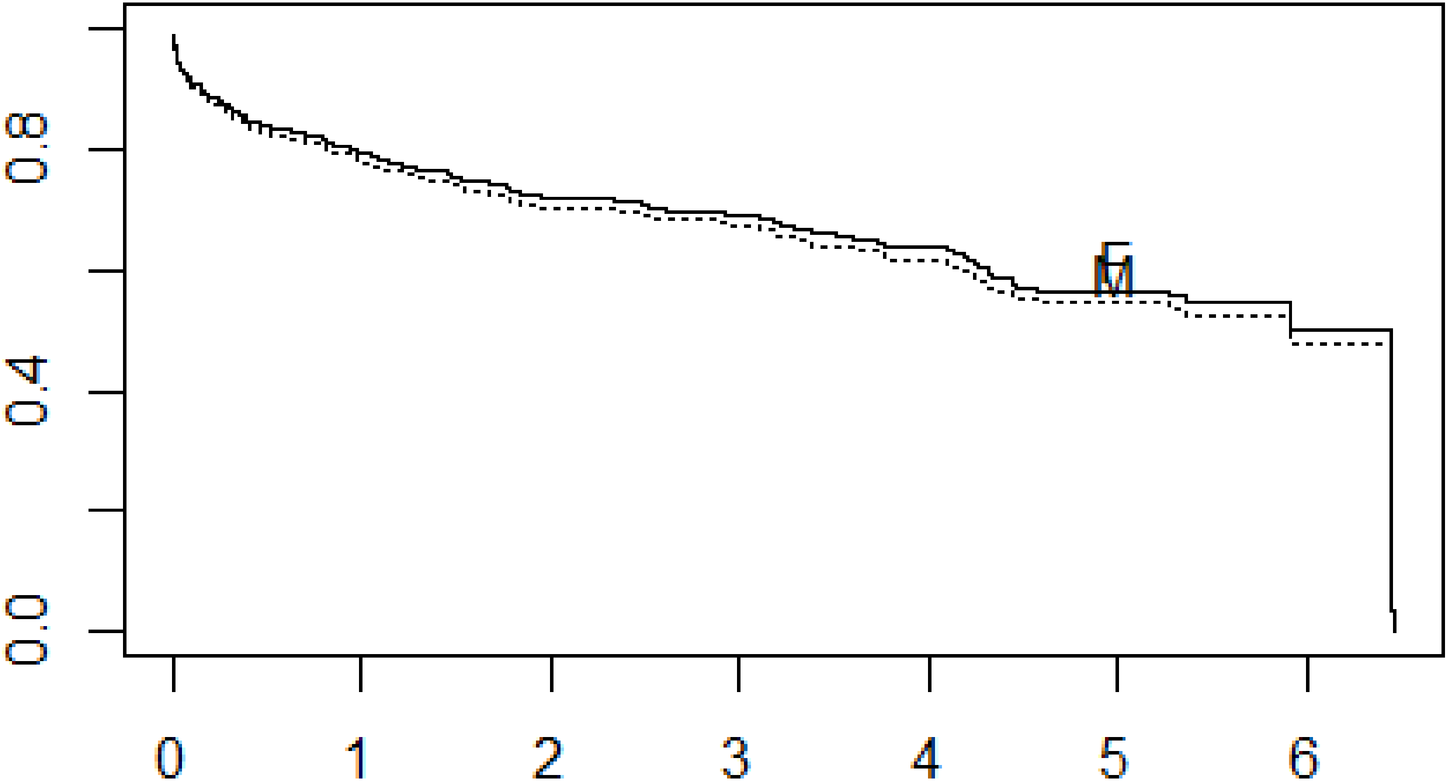
Adjusting for Covariate Imbalance

There is a 8.1 year difference between the average ages of men and women. The hazard ratio for age, 1.069, can get extrapolated to a 8.1 year difference by exponentiating. That is $1.069^{8.1} = 1.72$ which is actually larger than the hazard ratio that we saw for the unadjusted model with just gender.

66.6 Year Male Versus 74.7 Year Female



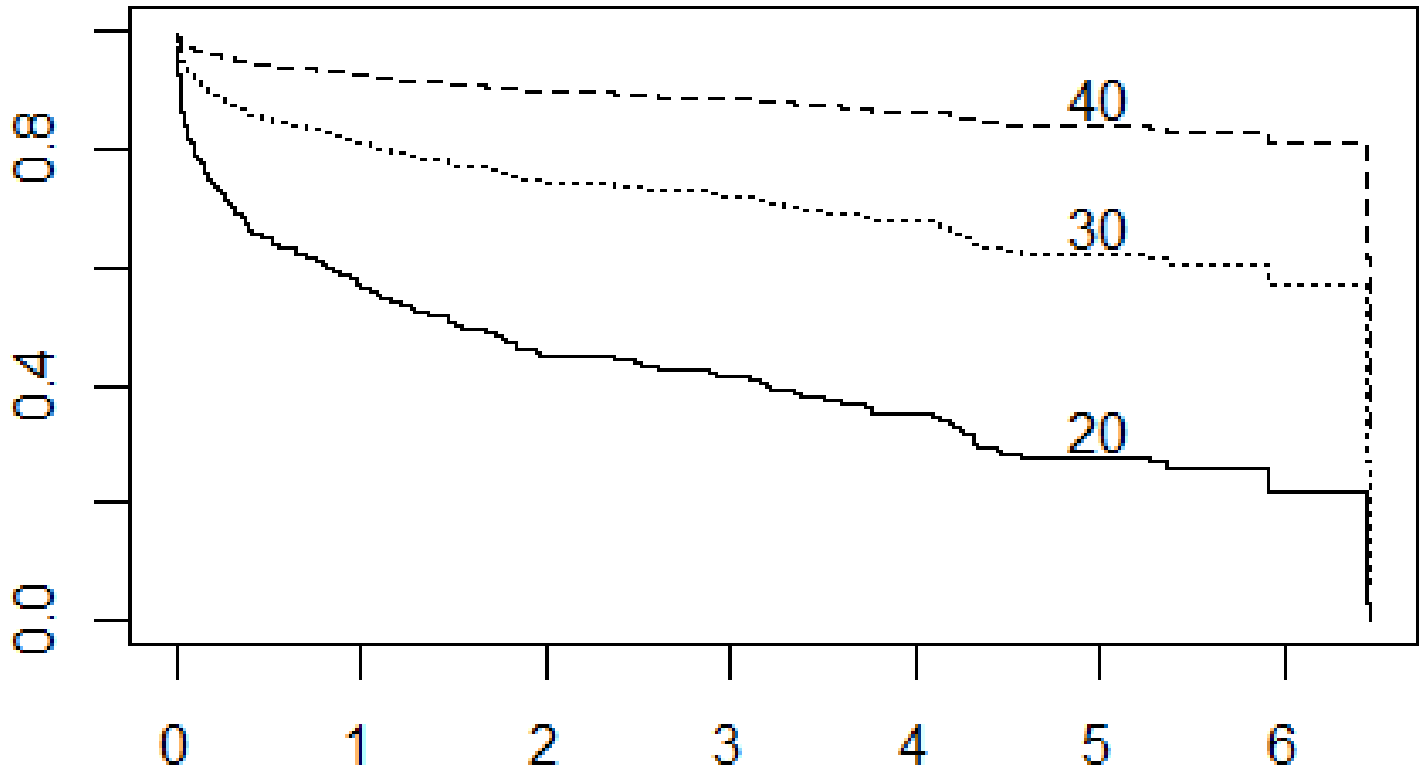
69.8 Year Male Versus 69.8 Year Female



Univariate Analysis of BMI

```
##   term   hr p.value     conf.int
## 1  bmi 0.91   0.001 0.88 to 0.93
```

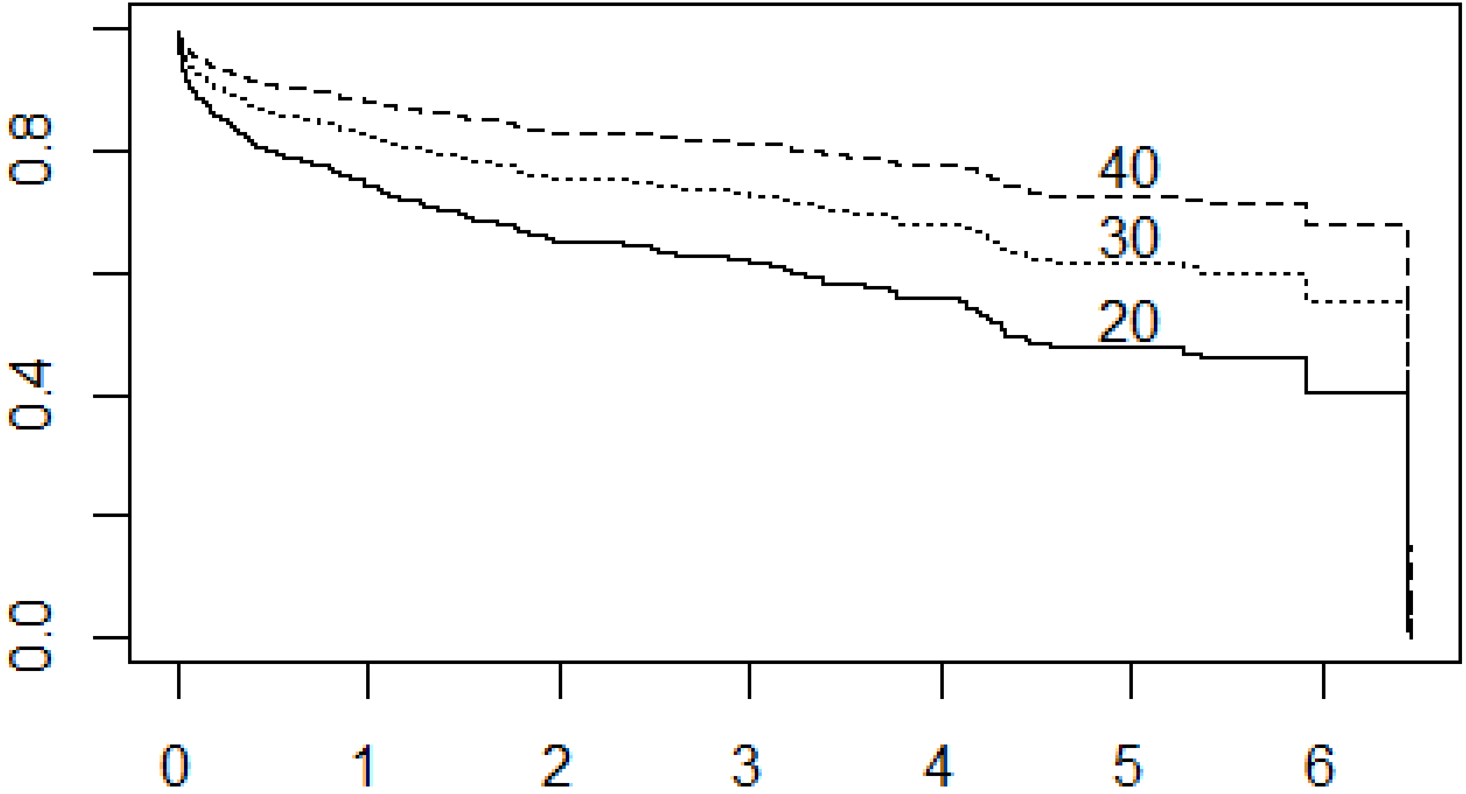
Unadjusted Survival Curves for Different BMI Values



Adjusting BMI for Age, Gender

```
##      term   hr p.value     conf.int
## 1     bmi 0.96   0.509 0.93 to 0.99
## 2     age 1.06   0.509 1.05 to 1.08
## 3 i_female 0.91   0.509  0.69 to 1.2
```

Adjusted BMI Survival Plots



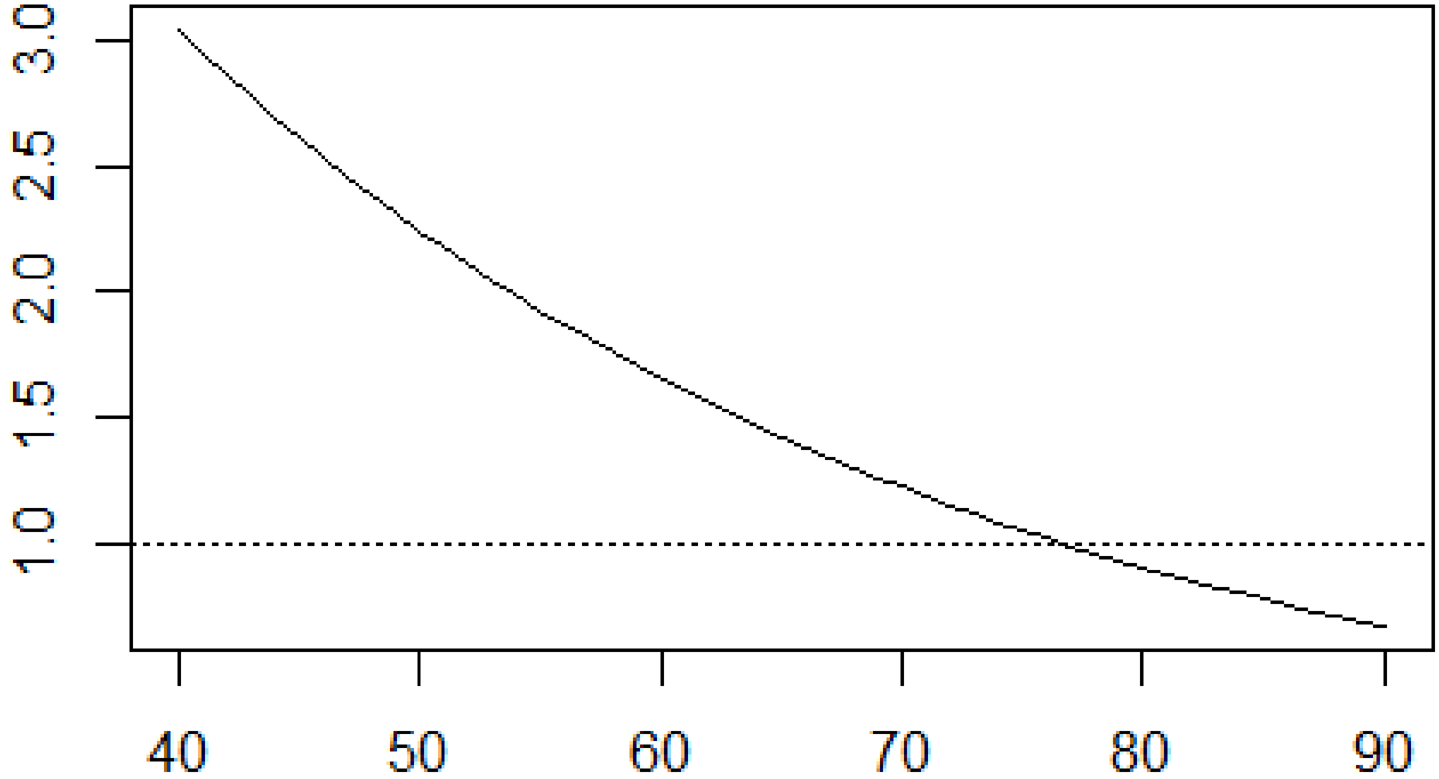
An Interaction Model; the Raw Interaction is Hard to Interpret

```
##            term    hr p.value       conf.int
## 1           age  1.08   0.019   1.06 to 1.1
## 2      i_female 10.32   0.019 1.47 to 72.19
## 3 age:i_female  0.97   0.019  0.95 to 0.99
```

Interaction Using Centered Values is Easier to Interpret

```
##              term   hr p.value     conf.int
## 1           age_c 1.08   0.244  1.06 to 1.1
## 2        i_female 1.23   0.244 0.87 to 1.73
## 3 age_c:i_female 0.97   0.244 0.95 to 0.99
```

Gender Hazard Ratio by Age



You Can Use a Sequence of Wald Tests to Compare Different Models

```
## Model 1
##          term   hr p.value     conf.int
## 1 genderFemale 1.46   0.006 1.12 to 1.92
##
## Model 2
##          term   hr p.value     conf.int
## 1          age 1.07   0.637 1.06 to 1.08
## 2 genderFemale 0.94   0.637 0.71 to 1.23
##
## Model 3
##          term   hr p.value     conf.int
## 1          bmi 0.96   0.509 0.93 to 0.99
## 2          age 1.06   0.509 1.05 to 1.08
## 3 genderFemale 0.91   0.509  0.69 to 1.2
```

Comparing Using Likelihoods

You use the log partial likelihood and/or the AIC (Akaike Information Criteria) to compare models of different complexity.

AIC = -2 log Likelihood + 2 k.

AIC = -2 log Likelihood + log(n) k.

AIC Comparisons

```
##                lab    logLik      AIC      BIC
## 1       gender only -1223.522 2449.043 2452.414
## 2       gender, age -1156.138 2316.276 2323.017
## 3 gender, age, bmi -1152.310 2310.620 2320.732
```

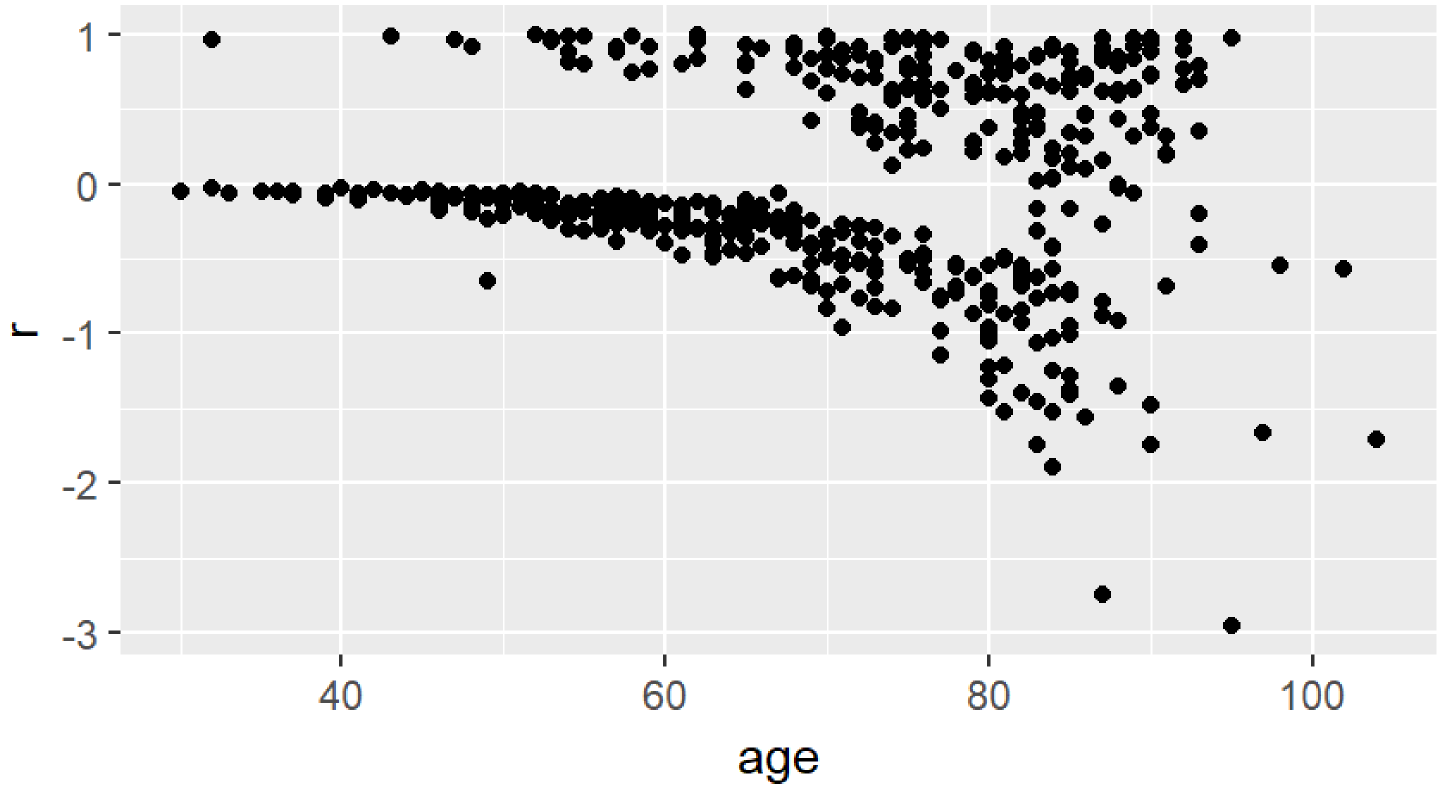
Martingale Residuals

There are several residuals available for Cox regression. The Martingale residual is defined as

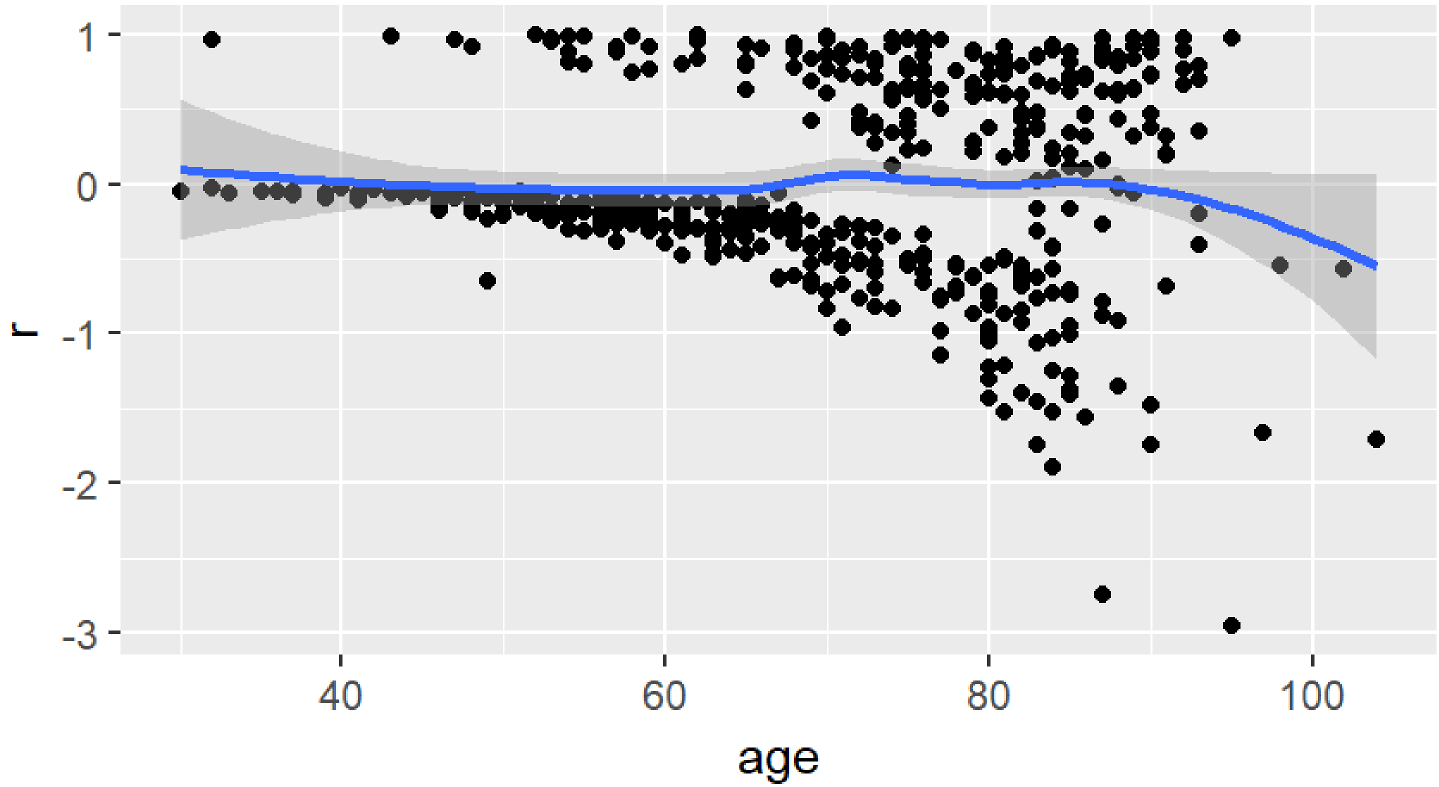
$$M(t_i) = \delta_i - H_0(t_i) e^{X\beta}$$

where δ_i = 0 if censored, 1 if dead.

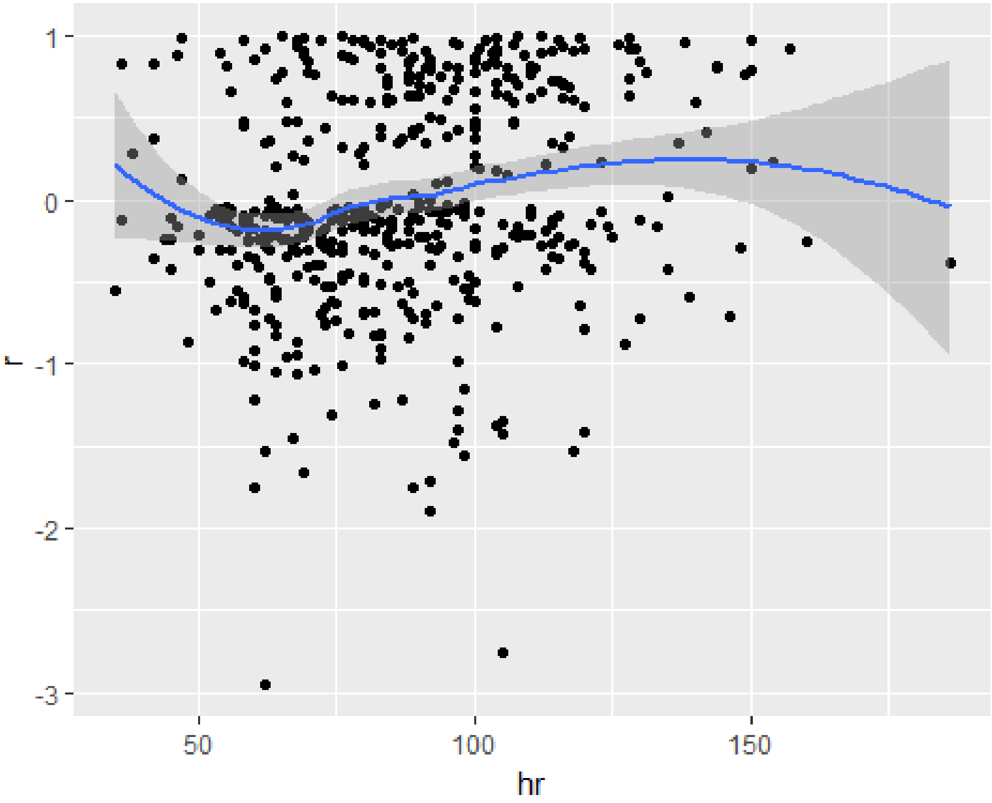
Residual Plot for Age



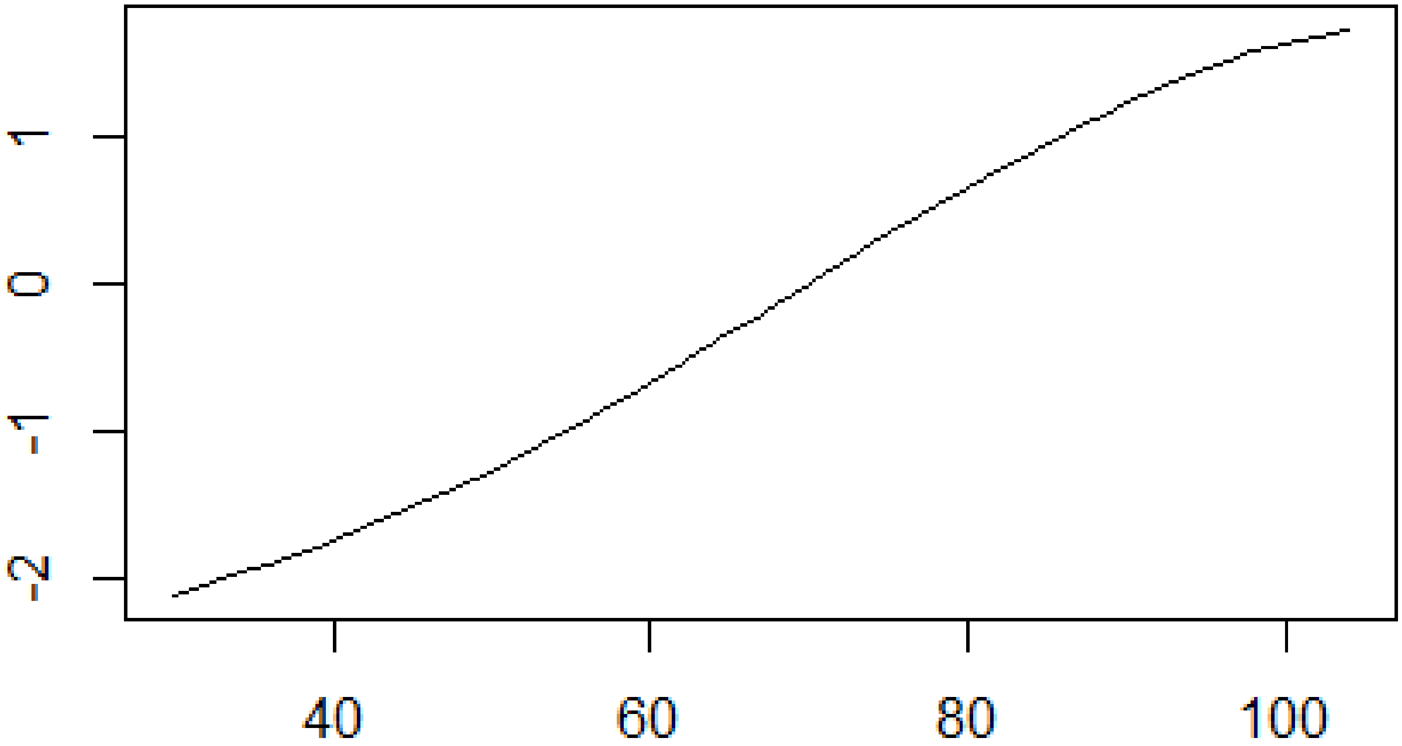
Residual Plot for Age with Smoothing Line



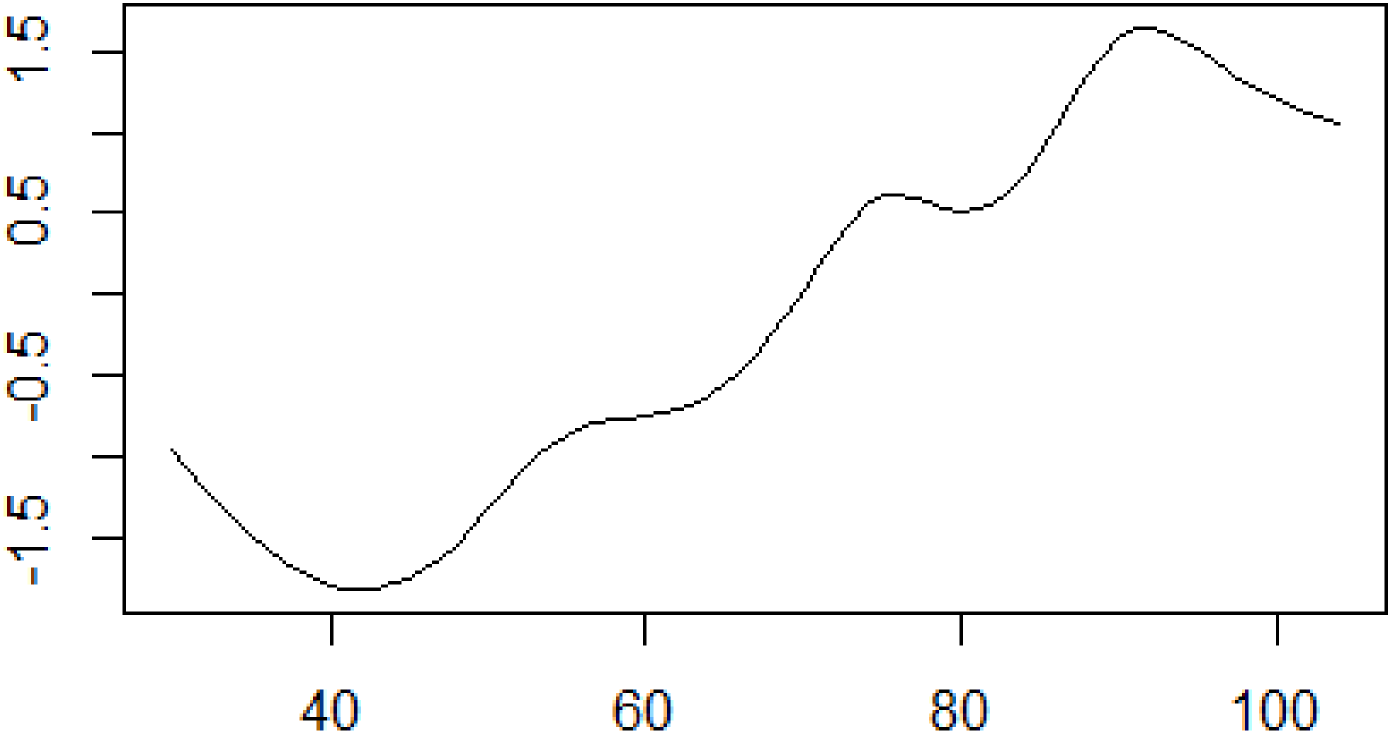
Residual Plots for HR



Using Splines to Model Non-linearities



An 8 df Spline (Overfitting!)



Comparing Linear Model to Two Splines

```
##              lab    logLik      AIC      BIC
## 1 linear (df=1) -1152.310 2310.620 2320.732
## 2 spline (df=3) -1151.288 2308.714 2319.058
## 3 spline (df=8) -1143.118 2302.204 2329.116
```

Next Time - Testing Proportional Hazards

1. Patterns in Kaplan-Meier curves
2. Complementary log-log plot
3. Schoenfeld Residuals
4. Fit time varying covariates

What Have You Learned Today?

1. The Cox regression model allows for multiple independent variables and interactions.
2. The predicted survival curve estimated at a common covariate mean produces a risk-adjusted comparison.
3. A positive martingale residual implies a death earlier than expected.