Cox Regression

… and a review of the hazard function

Steve Simon

Abstract

When you have data measuring the time to an event, you can examine the relationship between various predictor variables and the time to the event using a Cox proportional hazards model. In this talk, you will see what a hazard function is and describe the interpretations of increasing, decreasing, and constant hazard. Then you will examine the log rank test, a simple test closely tied to the Kaplan-Meier curve, and the Cox proportional hazards model.

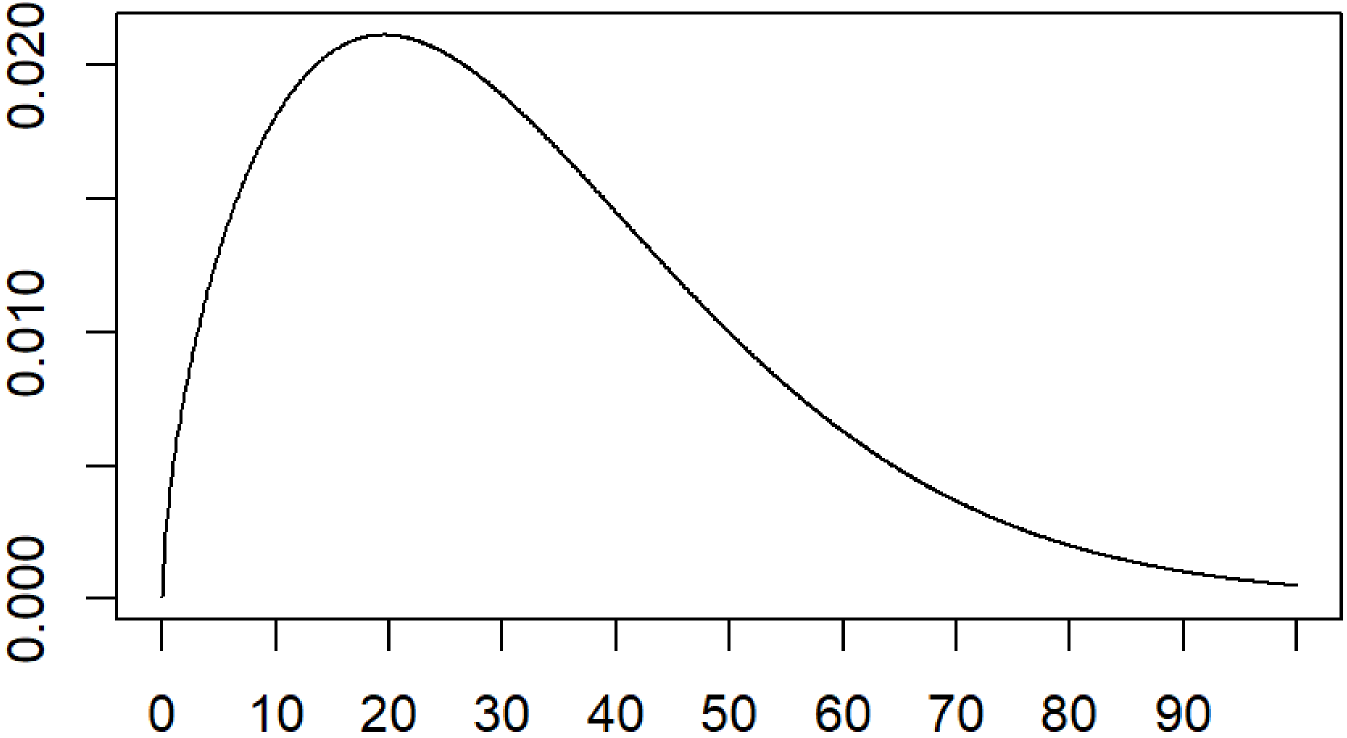
Defining the Hazard Function

Rates versus proportions

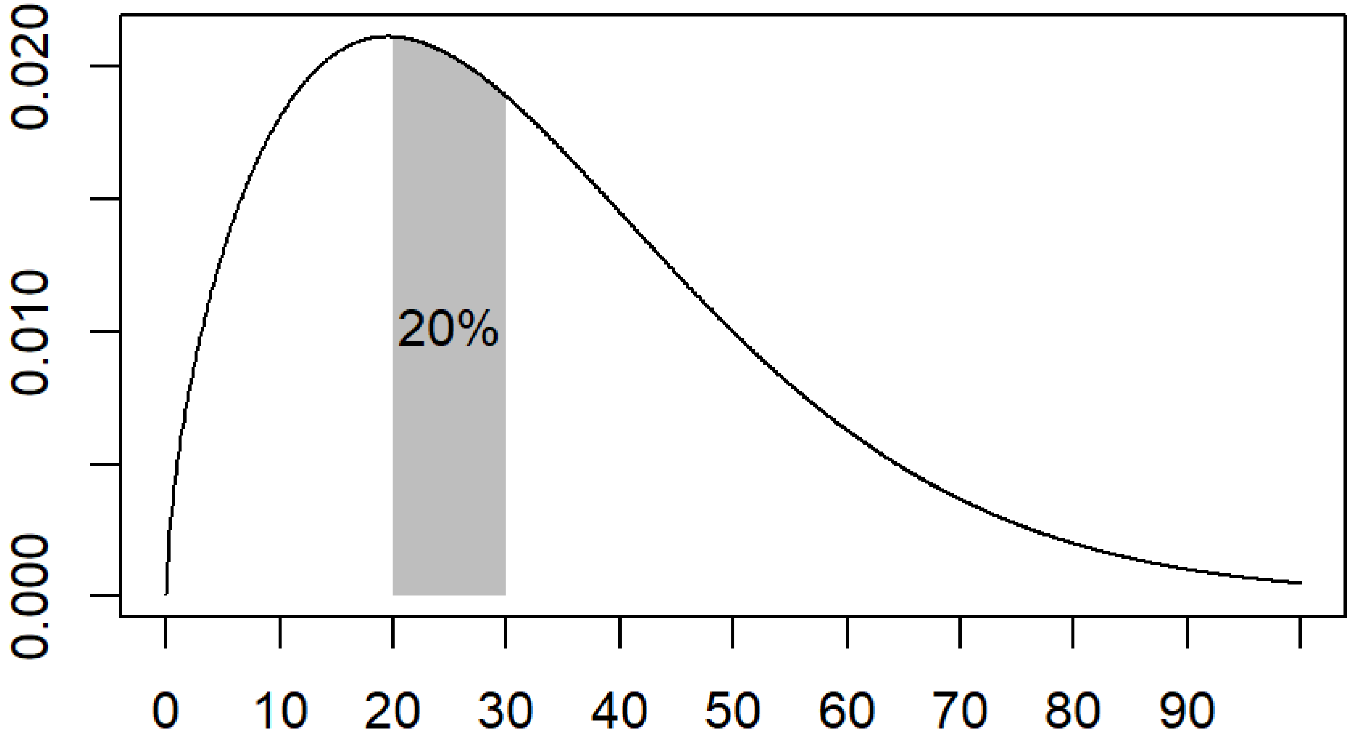
There is an important, but subtle distinction

- Both proportions and rates involve division.
- A proportion is a count divided by another count.
  - It can never be larger than 1.
- A rate is a count divided by a measure of time (or sometimes area).
  - It can sometimes be larger than 1.

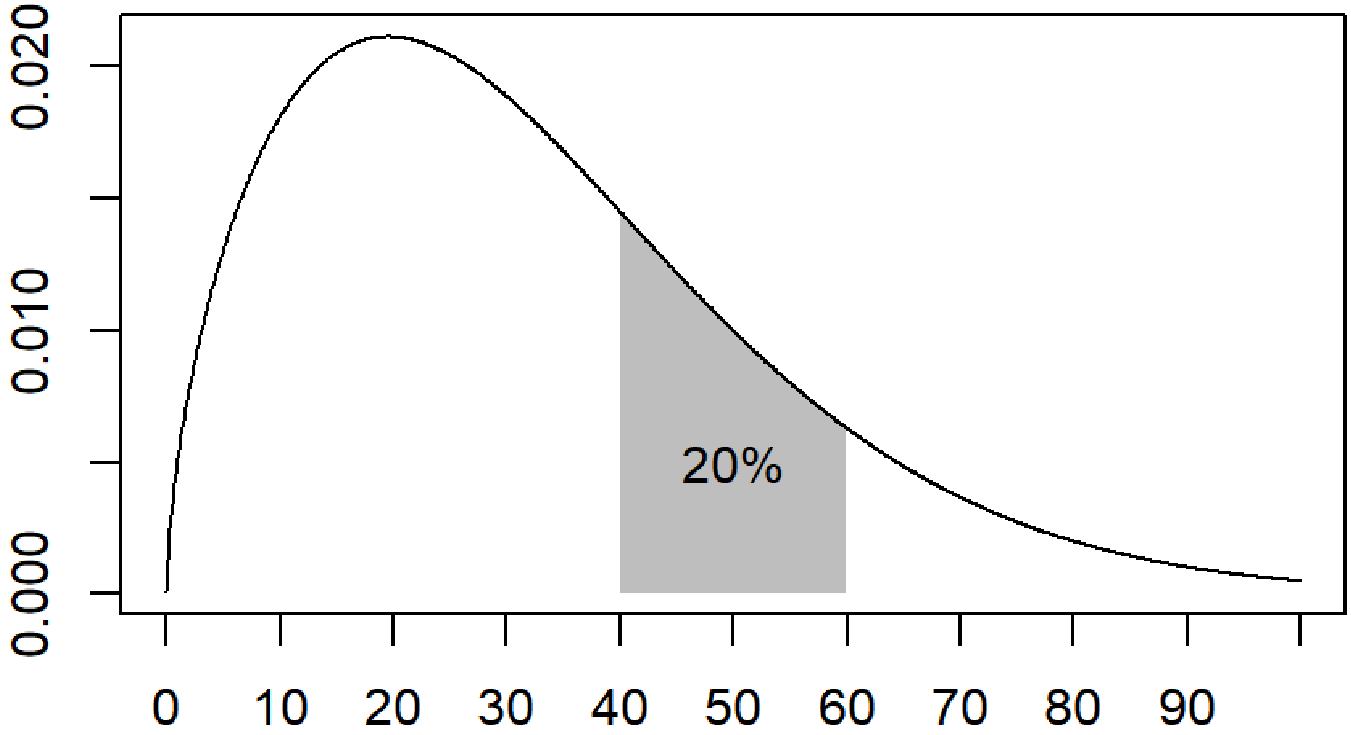
Defining the Hazard Function



Defining the Hazard Function



Defining the Hazard Function



Defining the Hazard Function

Are you comparing apples and oranges?

Three problems with these comparisons

1. The number of people alive at age 20 is much larger than the number of people alive at age 40.
2. The probabilities are measured across different time ranges.
3. The probabilities are quite heterogenous across the time intervals.

Defining the Hazard Function

To make a “fair” comparison

1. Adjust by the probability of surviving up to age 20 or age 40.
2. Calculate a death rate by dividing by the time range.
3. Calculate over a narrow time interval, Δt.

Defining the Hazard Function

Calculating a hazard function

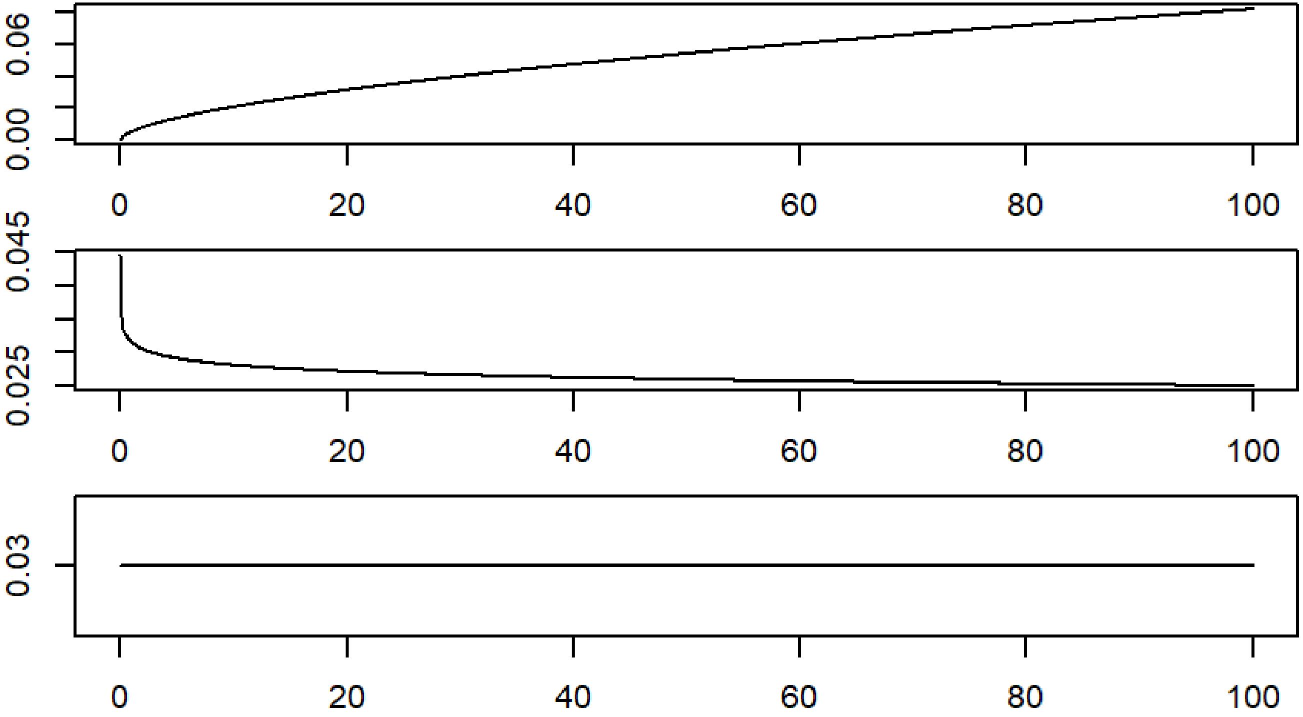
The hazard function is defined as

$$h(t) = (P[t \leq T \leq t+\Delta t] / \Delta t) / P[T \geq t]$$

Key points

1. adjusted for the number surviving to that time ($P[T \geq t]$),
2. calculated as a rate ($P[t \leq T \leq T+\Delta t] / \Delta t$), not a probability, and
3. computed over a narrow time interval.

Defining the Hazard Function



Why is the Hazard Function Important?

1. Knowledge of the hazard function for an industrial product can help you decide on how and when to employ preventive maintenance.
2. It can also help you extrapolate survival patterns beyond the range of observed values, such as for long-term reliability testing.
3. If you are planning a new research study, the hazard function can help you decide how many patients to study and the length of follow-up for these patients.
4. Assumptions about the hazard function are critical for regression models of survival.

The Cumulative Hazard Function

The cumulative hazard function, H(t), is defined as

$$H(t) = \int_0^t h(u)du$$

and it has an interesting relationship to the survival function.

$$S(t) = e^{-H(t)}$$

The Cumulative Hazard Function

Calculation of the hazard function, h(t), from a set of data is quite difficult. Fortunately, you can calculate the cumulative hazard function, H(t), without much difficulty.

$$\hat{H}(t_j) = \sum_{i=1}^{j} \frac{d_i}{n_i}$$

where d_i and n_i are the number of deaths and the number at risk at time t_i.

Recall the Limitations of the Log Rank Test

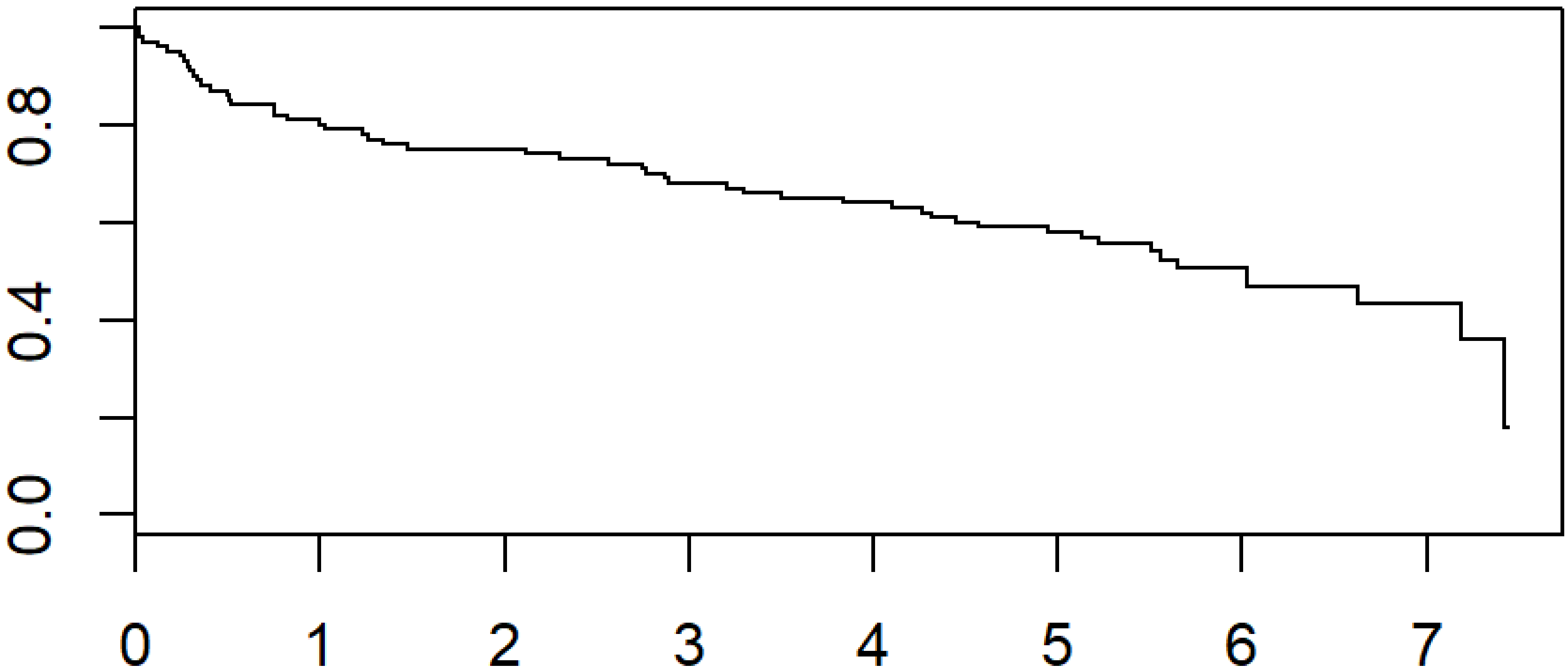
The log rank test

- works well when you’re comparing a treatment group to a control group
- you can also use it when you have three or more groups.

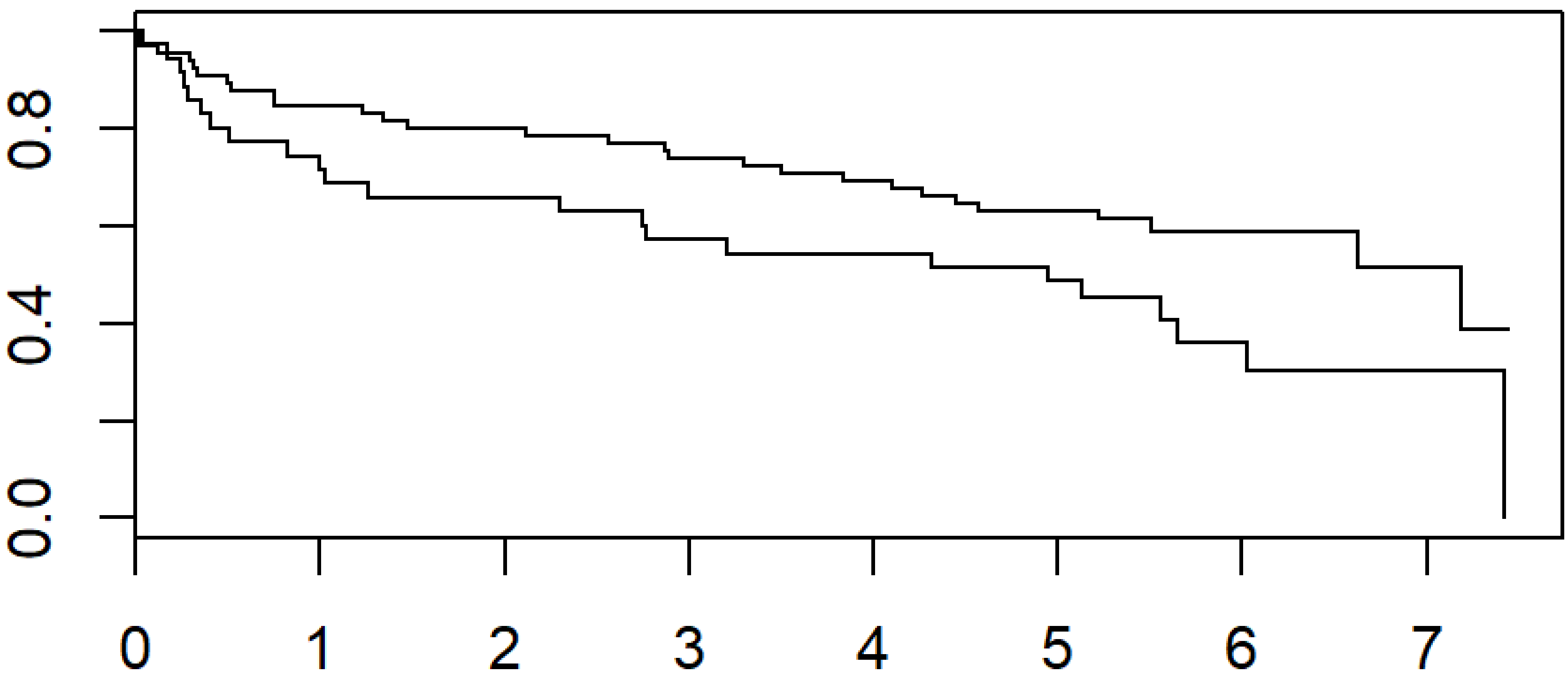
But the log rank test does not extend beyond this:

- you cannot include a continuous predictor,
- you cannot analyze data with multiple predictors, and
- you cannot do risk adjustment.

Always Start with a Plot of Overall Survival



Then a Kaplan-Meier Curve by Groups



The Log Rank Test

```
Call:
survdiff(formula = Surv(time_yrs, fstat == "Dead") ~ gender, 
    data = whas100)

               N Observed Expected (O-E)^2/E (O-E)^2/V
gender=Male    65       28     34.6      1.27      3.97
gender=Female  35       23     16.4      2.68      3.97

 Chisq= 4  on 1 degrees of freedom, p= 0.0463
```

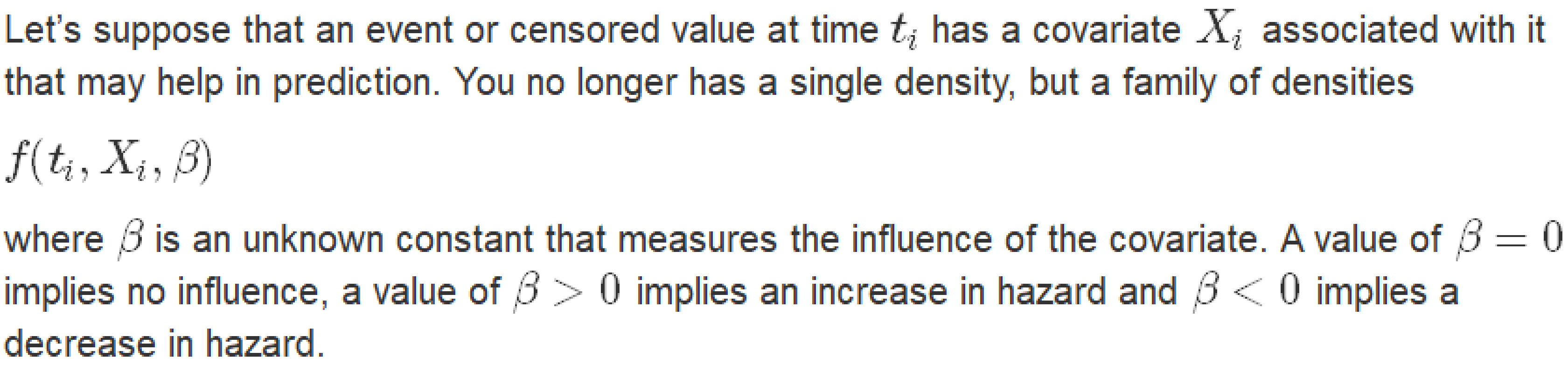
The Cox Regression Model

```
Call:
coxph(formula = Surv(time_yrs, whas100$fstat == "Dead") ~ 
gender,
    data = whas100)

  n= 100, number of events= 51 

               coef exp(coef) se(coef)     z Pr(>|z|)  
genderFemale 0.5548    1.7416   0.2824 1.965   0.0494 *
```

The Cox Regression Model

```
 exp(coef) exp(-coef) lower .95 upper .95
genderFemale     1.742     0.5742     1.001     3.029

Concordance= 0.565  (se = 0.035 )
Rsquare= 0.037   (max possible= 0.985 )
Likelihood ratio test= 3.75  on 1 df,   p=0.05293
Wald test            = 3.86  on 1 df,   p=0.04945
Score (logrank) test = 3.96  on 1 df,   p=0.04665
```

A Full Likelihood Approach

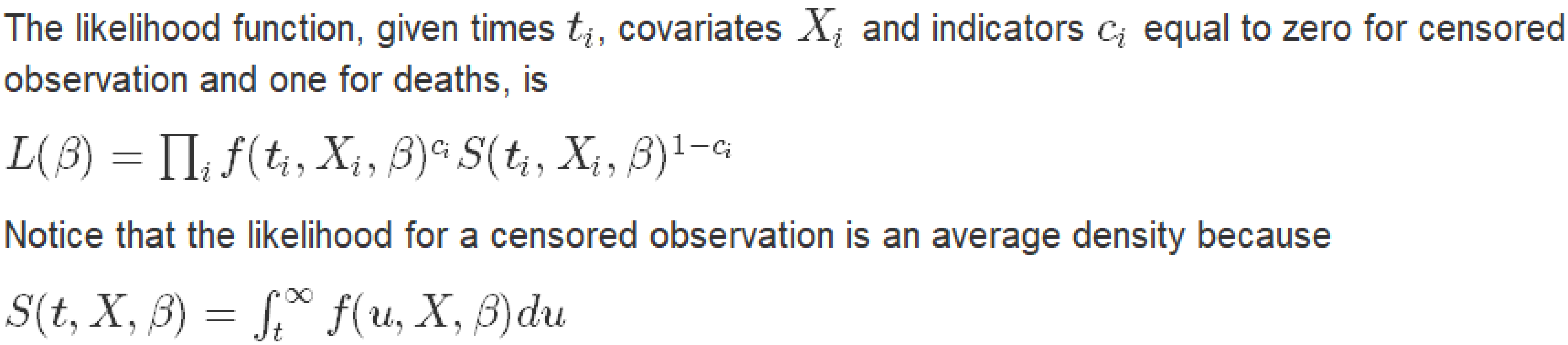


Let's suppose that an event or censored value at time t_i has a covariate X_i associated with it that may help in prediction. You no longer has a single density, but a family of densities

$$f(t_i, X_i, \beta)$$

where β is an unknown constant that measures the influence of the covariate. A value of $\beta = 0$ implies no influence, a value of $\beta > 0$ implies an increase in hazard and $\beta < 0$ implies a decrease in hazard.

A Full Likelihood Approach



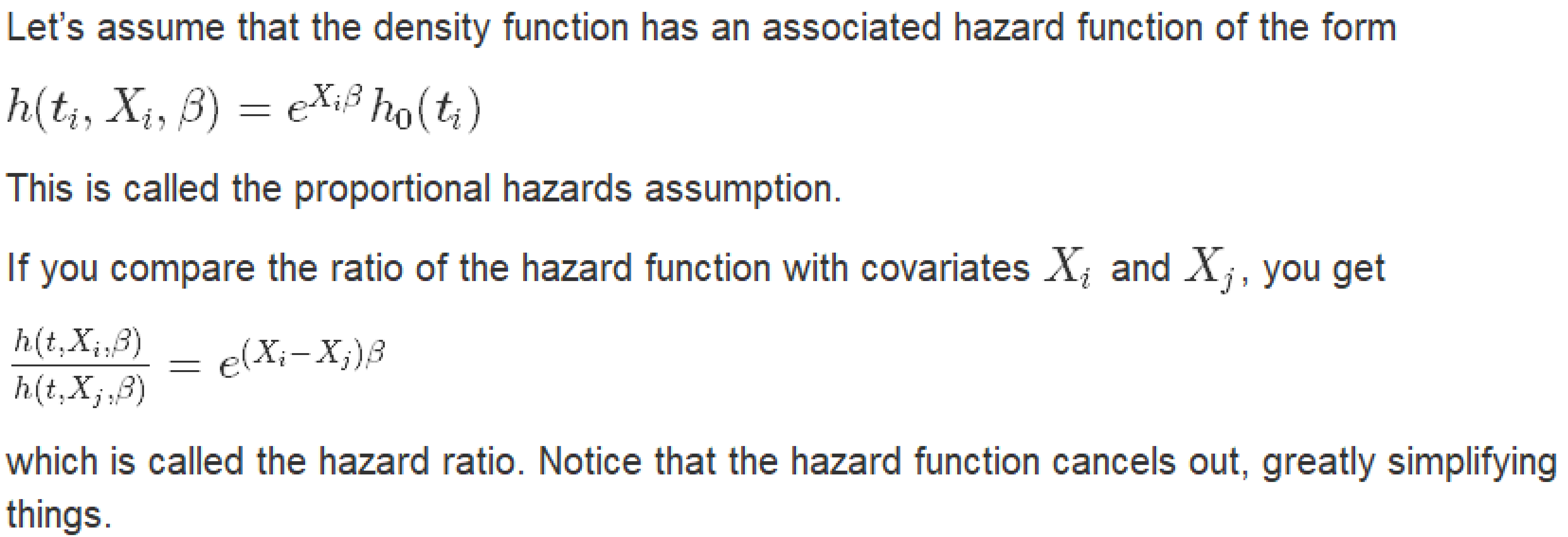
The likelihood function, given times t_i, covariates X_i and indicators c_i equal to zero for censored observation and one for deaths, is

$$L(\beta) = \prod_i f(t_i, X_i, \beta)^{c_i} S(t_i, X_i, \beta)^{1-c_i}$$

Notice that the likelihood for a censored observation is an average density because

$$S(t, X, \beta) = \int_t^\infty f(u, X, \beta) du$$

Proportional Hazards Assumption



Let’s assume that the density function has an associated hazard function of the form

$$h(t_i, X_i, \beta) = e^{X_i\beta} h_0(t_i)$$

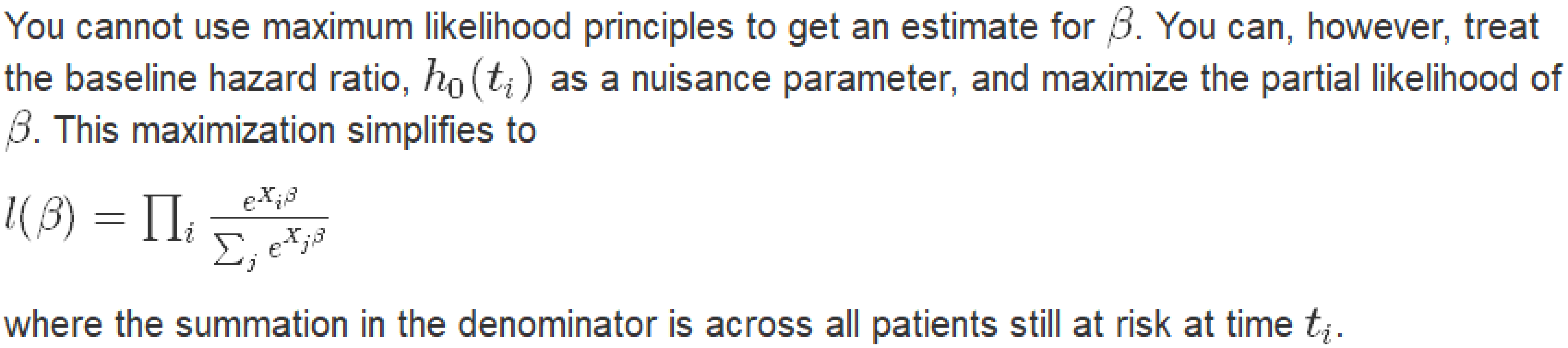
This is called the proportional hazards assumption.

If you compare the ratio of the hazard function with covariates X_i and X_j, you get

$$\frac{h(t, X_i, \beta)}{h(t, X_j, \beta)} = e^{(X_i - X_j)\beta}$$

which is called the hazard ratio. Notice that the hazard function cancels out, greatly simplifying things.

Estimation Via Partial Likelihood

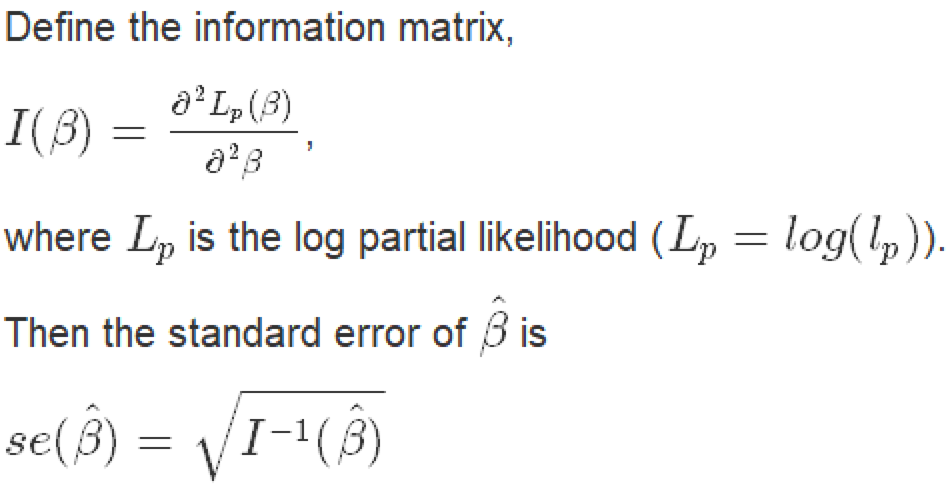


You cannot use maximum likelihood principles to get an estimate for β. You can, however, treat the baseline hazard ratio, $h_0(t_i)$ as a nuisance parameter, and maximize the partial likelihood of β. This maximization simplifies to

$$l(\beta) = \prod_i \frac{e^{X_i\beta}}{\sum_j e^{X_j\beta}}$$

where the summation in the denominator is across all patients still at risk at time t_i.

The Information Matrix



Define the information matrix,

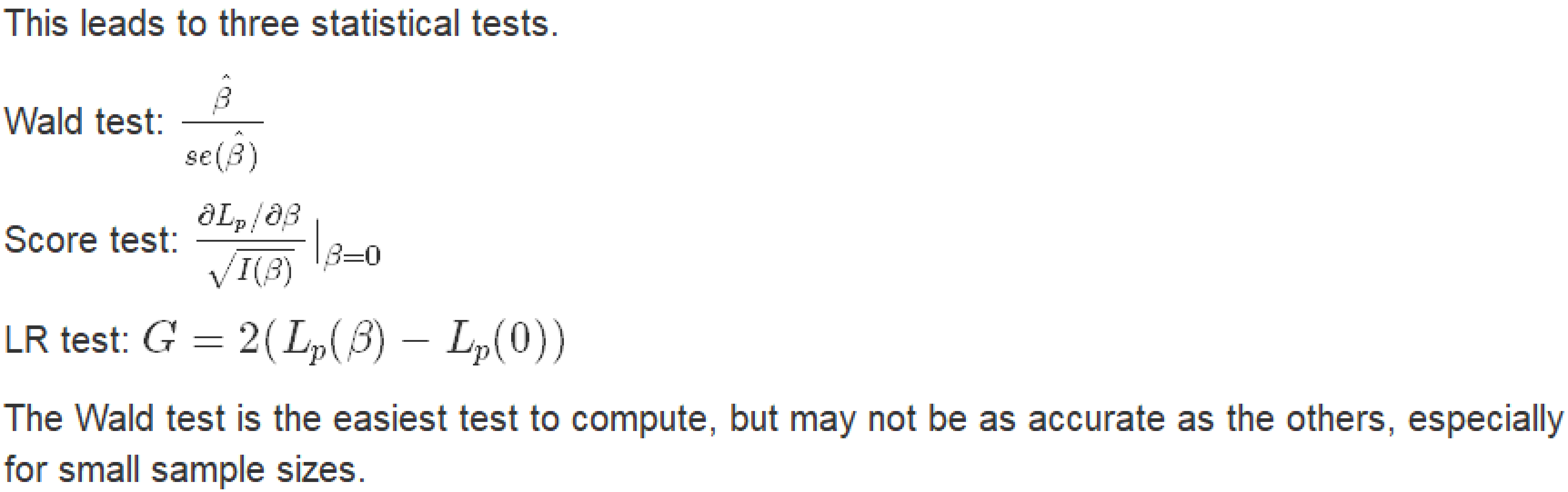
$$I(\beta) = \frac{\partial^2 L_p(\beta)}{\partial^2 \beta},$$

where L_p is the log partial likelihood ($L_p = log(l_p)$).

Then the standard error of $\hat{\beta}$ is

$$se(\hat{\beta}) = \sqrt{I^{-1}(\hat{\beta})}$$

The Wald, Score, and Likelihood Ratio Tests



This leads to three statistical tests.

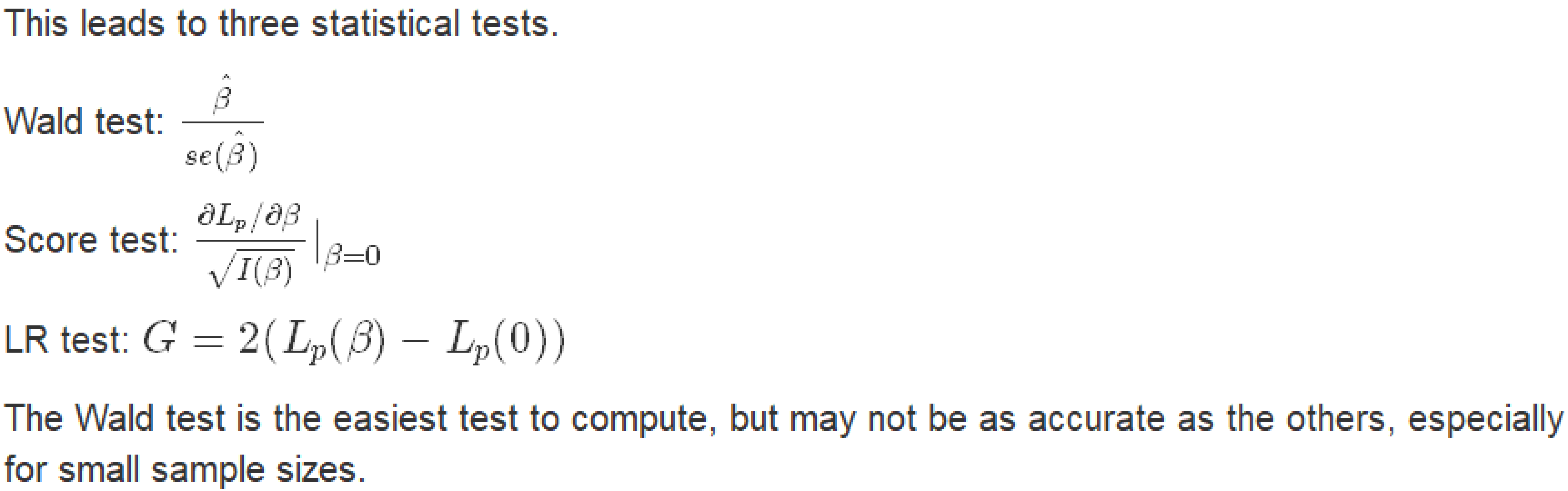
Wald test: $\frac{\hat{\beta}}{se(\hat{\beta})}$

Score test: $\frac{\partial L_p / \partial \beta}{\sqrt{I(\beta)}}\Big|_{\beta=0}$

LR test: $G = 2(L_p(\beta) - L_p(0))$

The Wald test is the easiest test to compute, but may not be as accurate as the others, especially for small sample sizes.

Revisiting the Cox Regression Model



This leads to three statistical tests.

Wald test: $\frac{\hat{\beta}}{se(\hat{\beta})}$

Score test: $\frac{\partial L_p / \partial \beta}{\sqrt{I(\beta)}} \Big|_{\beta=0}$

LR test: $G = 2(L_p(\beta) - L_p(0))$

The Wald test is the easiest test to compute, but may not be as accurate as the others, especially for small sample sizes.

Revisiting the Cox Regression Model

```
             exp(coef) exp(-coef) lower .95 upper .95
genderFemale     1.742     0.5742     1.001     3.029
```

Since you used an indicator variable for gender, the exponential of the coefficient, 1.742, is the hazard ratio. Women are at 74% greater risk of mortality in this data set than men.

Revisiting the Cox Regression Model

```
 coef exp(coef) se(coef)      z Pr(>|z|)
genderFemale 0.5548    1.7416   0.2824 1.965   0.0494 *

                    exp(coef) exp(-coef) lower .95 upper .95
genderFemale        1.742      0.5742     1.001     3.029
```

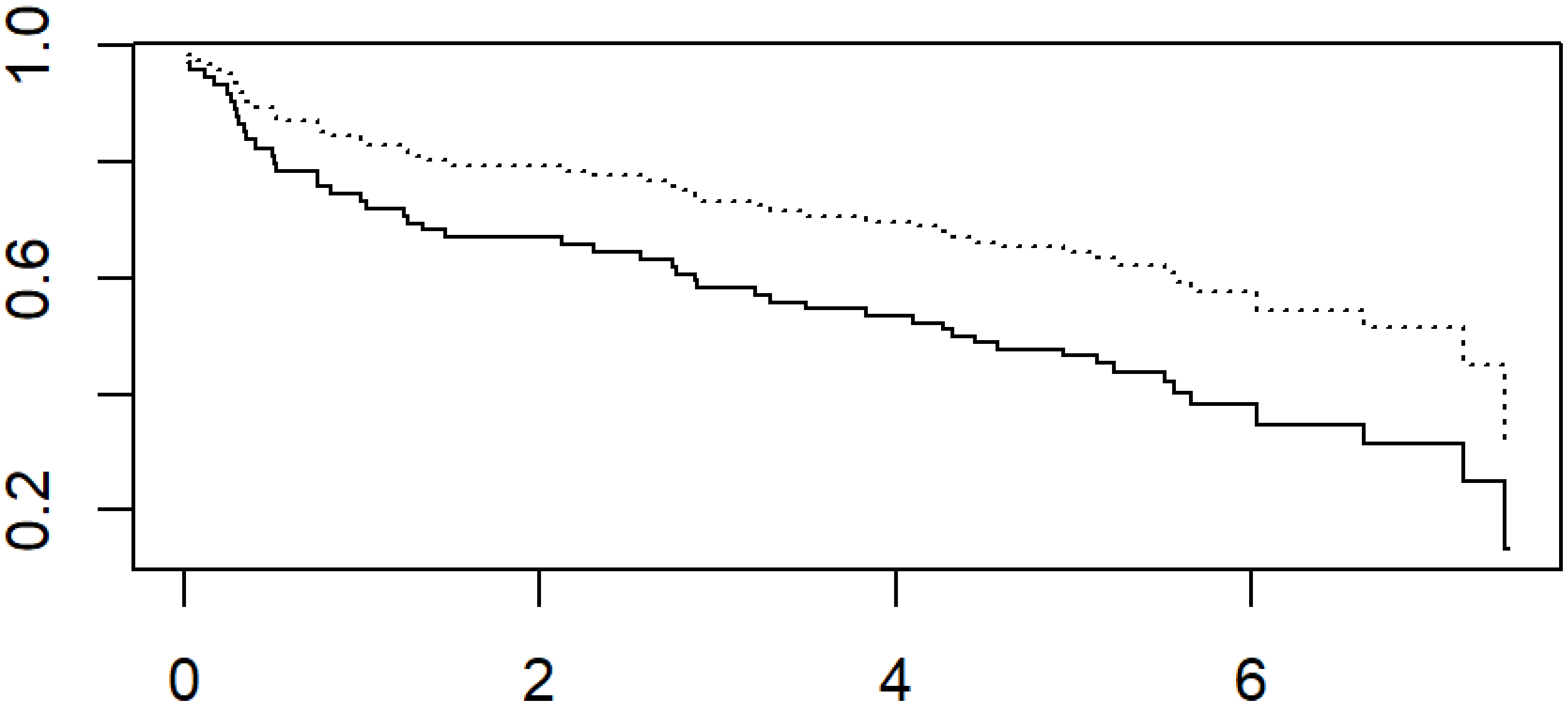
The 95% confidence interval for the hazard ratio is exp(0.5548 +/- 1.96*0.2824) = 1.001 to 3.029.

Revisiting the Cox Regression Model

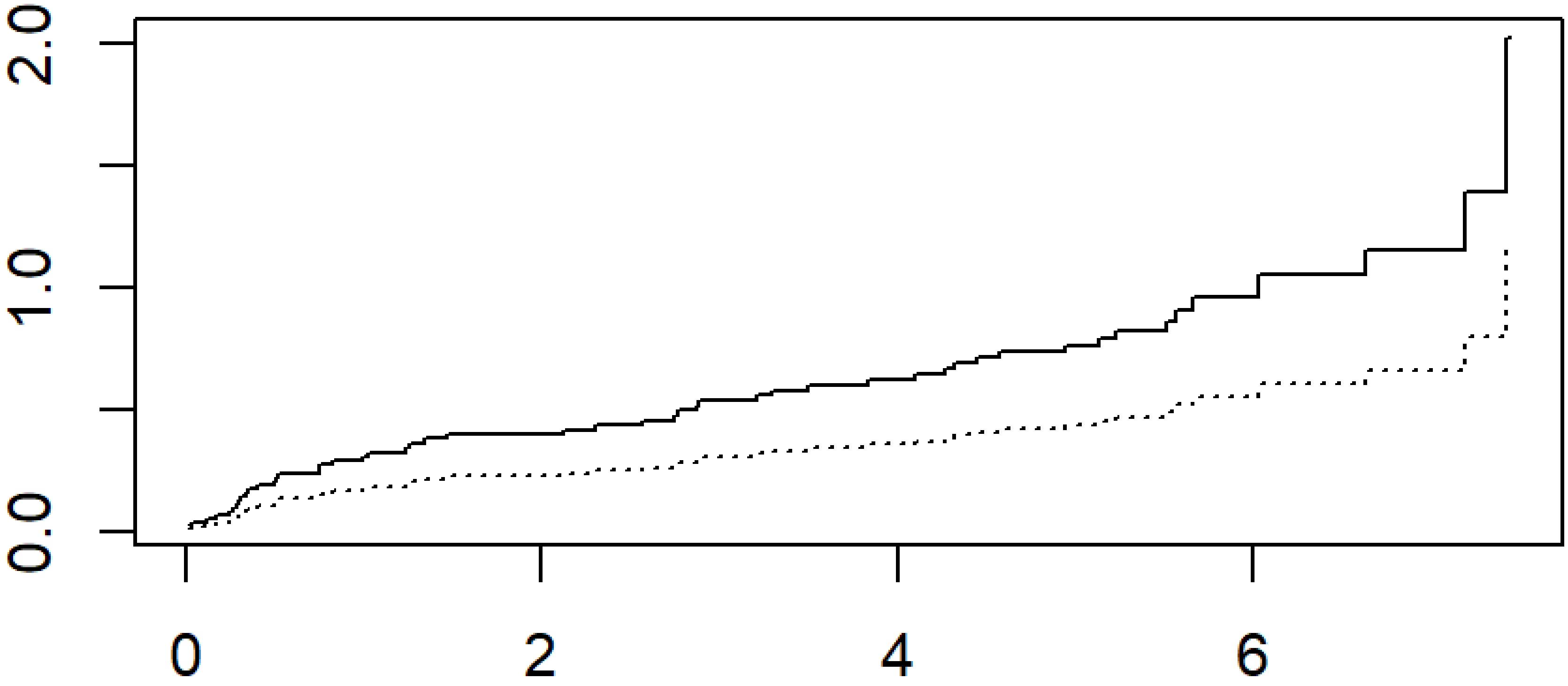
```
Likelihood ratio test= 3.75  on 1 df,   p=0.05293
Wald test            = 3.86  on 1 df,   p=0.04945
Score (logrank) test = 3.96  on 1 df,   p=0.04665
```

The three different statistical tests all produce (more or less) the same result. Note that the Wald test here is a square of the Z value of 1.965, and you should compare this to a chi-squared distribution.

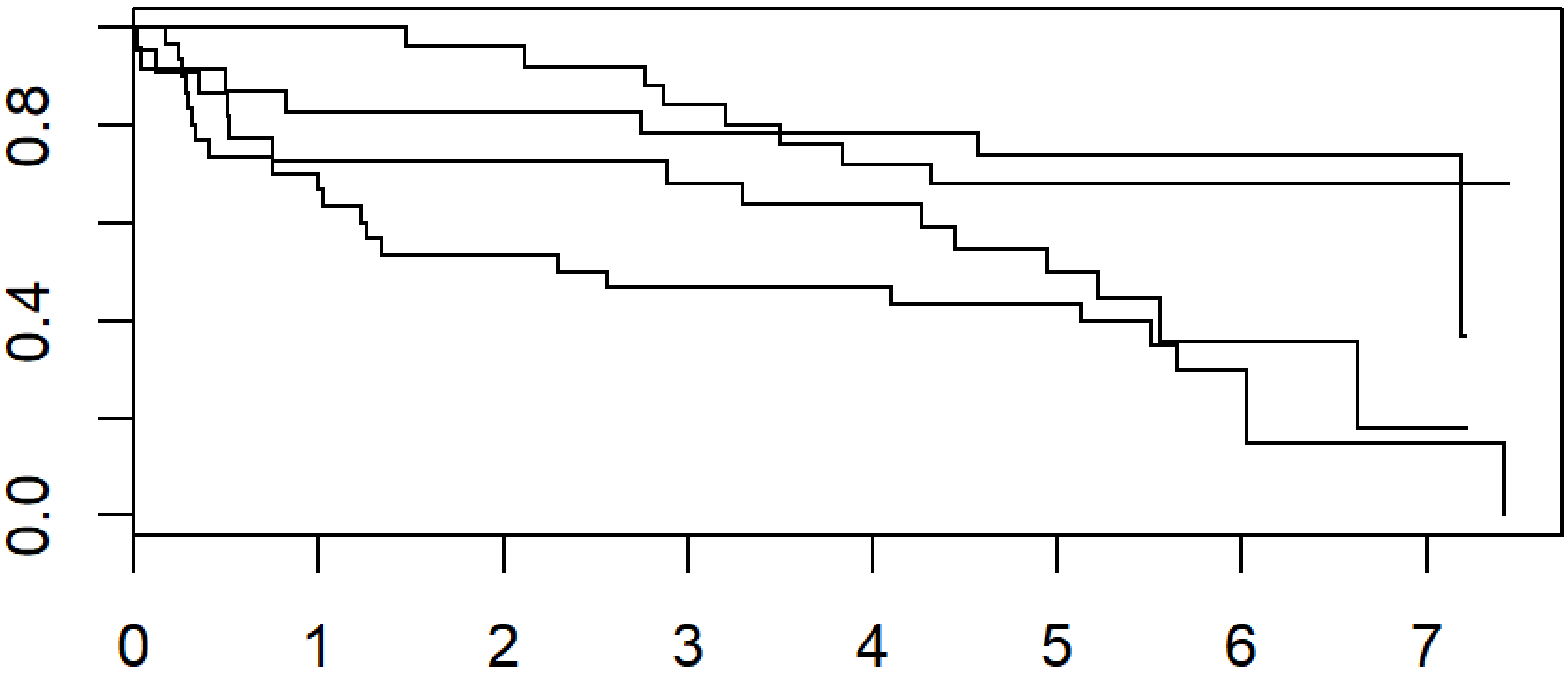
You Can Plot the Two Estimated Survival Curves



..and the Estimated Cumulative Hazard Functions



Start with the Basic Kaplan-Meier Curves



…and the Logrank Test

```
                    N Observed Expected (O-E)^2/E (O-E)^2/V
age_group=<60    25        8     15.5      3.64      5.29
age_group=60-69 23        7     12.9      2.71      3.68
age_group=70-79 22       14     10.2      1.39      1.76
age_group=>=80   30       22     12.3      7.57     10.18

 Chisq= 15.6  on 3 degrees of freedom, p= 0.00139
```

Here are the Cox Regression Model Results

```
 coef exp(coef) se(coef)    z      p
age_group60-69 0.0472    1.0484   0.5186 0.09 0.9274
age_group70-79 0.9866    2.6820   0.4454 2.22 0.0267
age_group>=80  1.2634    3.5373   0.4155 3.04 0.0024

Likelihood ratio test=15.3  on 3 df, p=0.00155
n= 100, number of events= 5139
```

Here are the Cox Regression Model Results

```
                exp(coef) exp(-coef) lower .95 upper .95
age_group60-69      1.048     0.9539    0.3794     2.897
age_group70-79      2.682     0.3729    1.1204     6.420
age_group>=80       3.537     0.2827    1.5666     7.987
```

But You Can Include Age as a Continuous Covariate

```
    coef exp(coef) se(coef)     z Pr(>|z|)
age 0.04567   1.04673  0.01195 3.822 0.000132 ***

    exp(coef) exp(-coef) lower .95 upper .95
age     1.047     0.9554     1.022     1.072
```

```
Each additional year in age increases the risk of death by 4.7%.

Each additional decade increase in age leads to a hazard ratio of
1.047^10 = 1.579.
```

What Have You Learned Today?

1. The hazard function is the short term death rate among those patients surviving to a specific time.
2. The log rank test is a simple test for comparing two or more Kaplan-Meier curves.
3. The Cox regression model allows for continuous predictors and risk adjustment.
4. The Cox model assumes proportional hazard functions and compares groups using a hazard ratio.