Survival Analysis

Stat 550 Regression

Miles Keppler

Department of Statistical Science, University of Idaho

May 14, 2021

Roadmap

- What is survival analysis?
  - Censoring
- Components of survivor models and data
  - Exponential survival times
  - Simulating data
  - Visualizing the data
  - Visualizing the theoretical models
- Estimation and comparison
  - Kaplan-Meier and non-parametric tests
  - Parametric regression
  - Cox regression
- References

Survival Analysis

Model the rate of survivors against a dichotommous event. Events are often death, sickness or recovery, but the tools are not limited to biological populations and events. A good example of data we would perform survival analysis on is a 10 year study of individuals who experience a heart attack, to track their time until they experience another. We often use data from randomized clinical trials and cohort studies.

Censoring

A common issue with survival analysis is that individuals will leave the study for a reason that is not the event occuring, this created right-censored data. The 10 year study can end without a subject experiencing a heart attack, they could die of unrelated causes, or they could drop out of the study for another reason. The data we recieved from them is important though, so we can't just throw it out; we need to use it. We assume that censoring is independent of our events of interest, and typically are uninterested in the rate of censoring once we get into survival analysis.

Survival Times

The survival time of an individual is the time until an event occurs. These events are typically of a nature where we would be interested in only the first occurence, such as death or disease onset. The survival times of a group of individuals will have some density function $f(t)$ which represents the probability of an event at exactly time t.

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}.$$

For exponentially distributed survival times, this would be the pdf

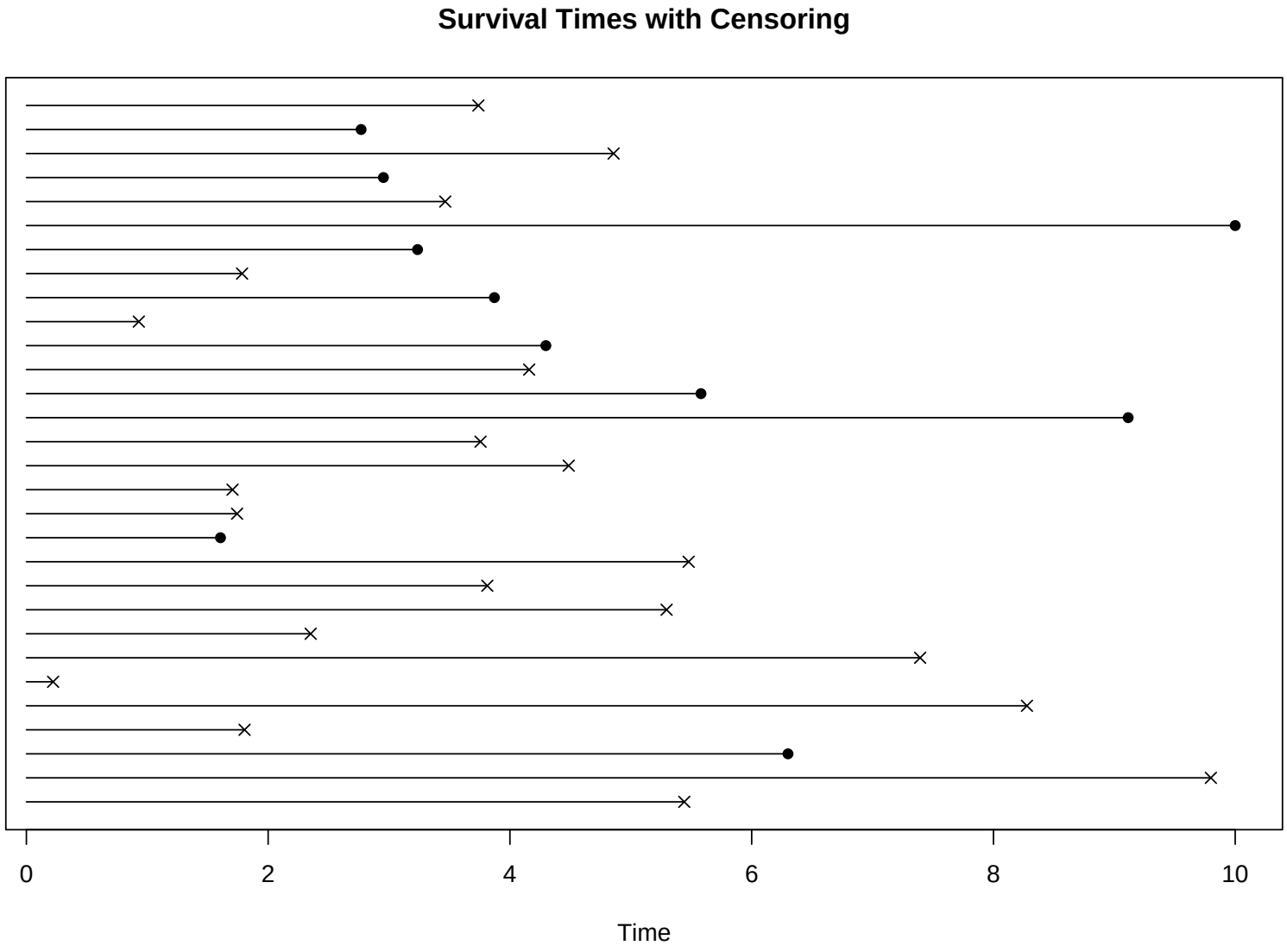
$$f(t) = \lambda e^{-\lambda x}.$$

Simulating Data

```
n = 30
lambda = .2
lifetime = rexp(n, rate = lambda)
censor = pmin(1 + rexp(n, rate = .1), 10)
endtime = pmin(lifetime, censor)
death = censor > lifetime
data = cbind.data.frame(endtime, death)
head(data)
```

```
##     endtime death
## 1 5.4424523  TRUE
## 2 9.7987102  TRUE
## 3 6.3003953 FALSE
## 4 1.8047928  TRUE
## 5 8.2772480  TRUE
## 6 0.2211451  TRUE
```

Data Visualized



Survival Function

The survival function $S(t)$ is the primary function of interest in survival analysis. It can be directly derived from the cdf $F(t)$ and takes the form

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= 1 - P(T < t) \\ &= 1 - F(t). \end{aligned}$$

Then the survival function of the exponential distribution would be

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= 1 - (1 - e^{-\lambda t}) \\ &= e^{-\lambda t}. \end{aligned}$$

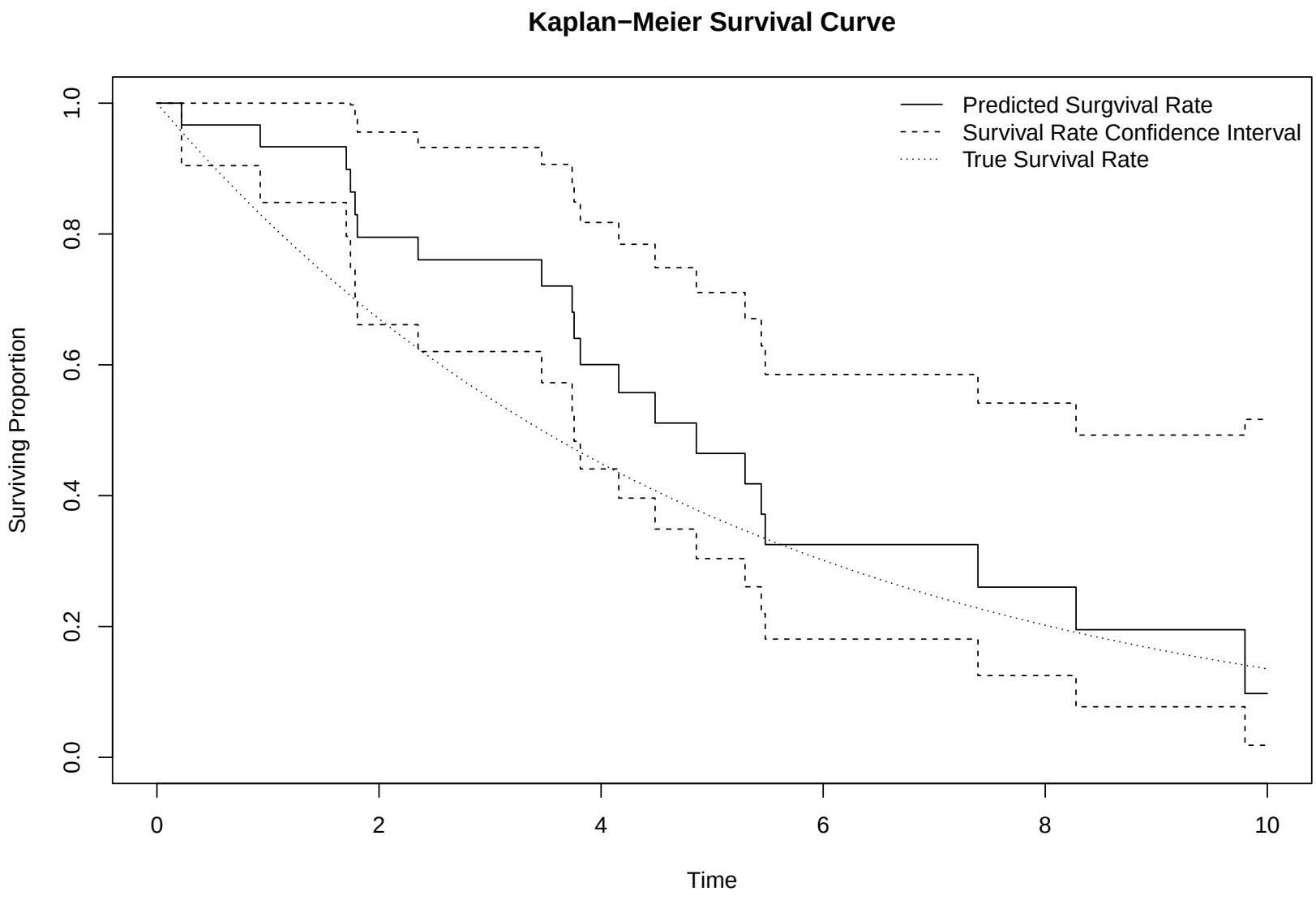
Kaplan-Meier Estimator

The Kaplan-Meier estimator $\hat{S}(t)$ is a non-parametric estimator for the survival function $S(t)$. The estimator is given by

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where d_i is the number of events at t_i and n_i is the number of known surviving individuals, not including censored individuals. This statistic is the non-parametric MLE of the survival rate at time t_i.

Survival Curve



Hazard Function

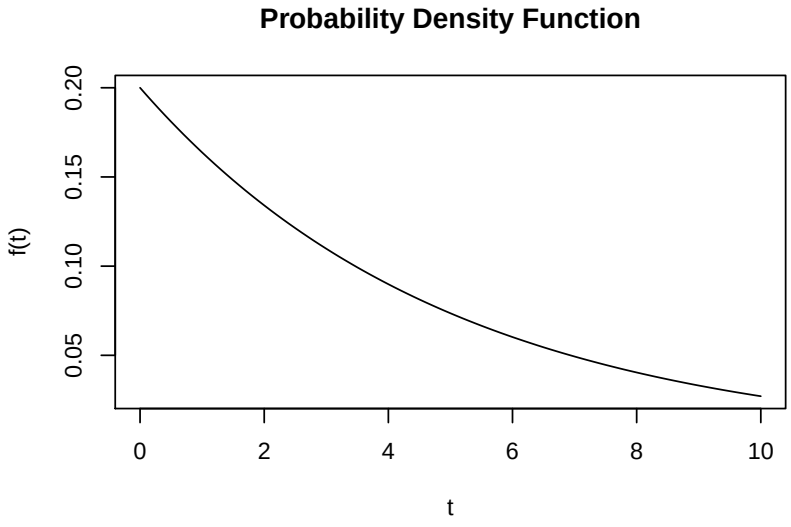
The hazard function $h(t)$ is the relative probability that an indivudual who has survived until time t will succumb immediately. This is essentially a relativee probability of survival given that the individual has survived until time t.

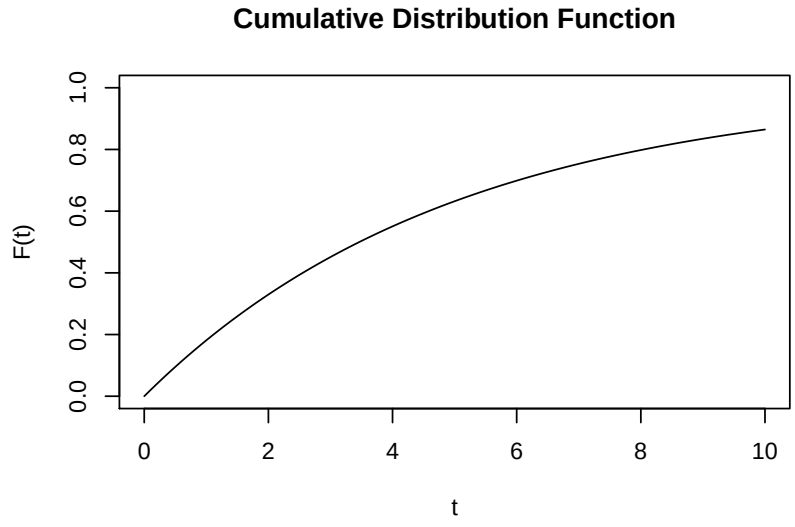
$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T \geq t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

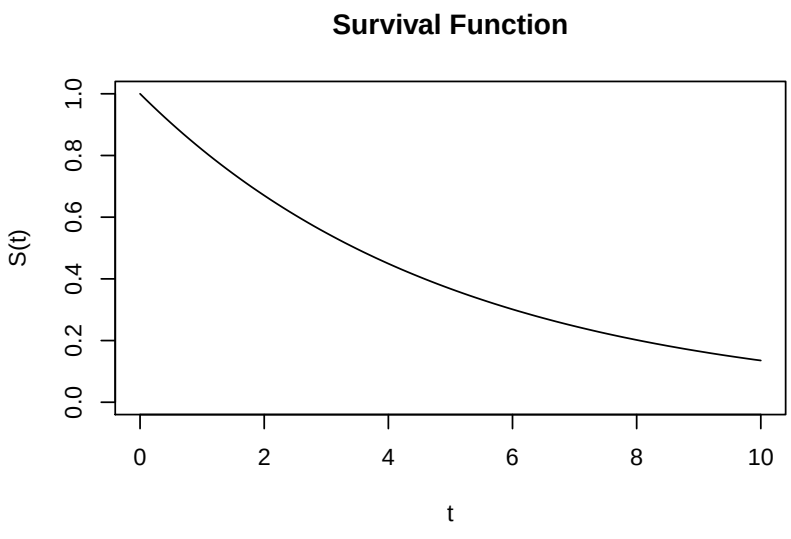
The hazard function $h(t)$ is constant for survival times derived from the exponential distribution, making it a particularly simple example.

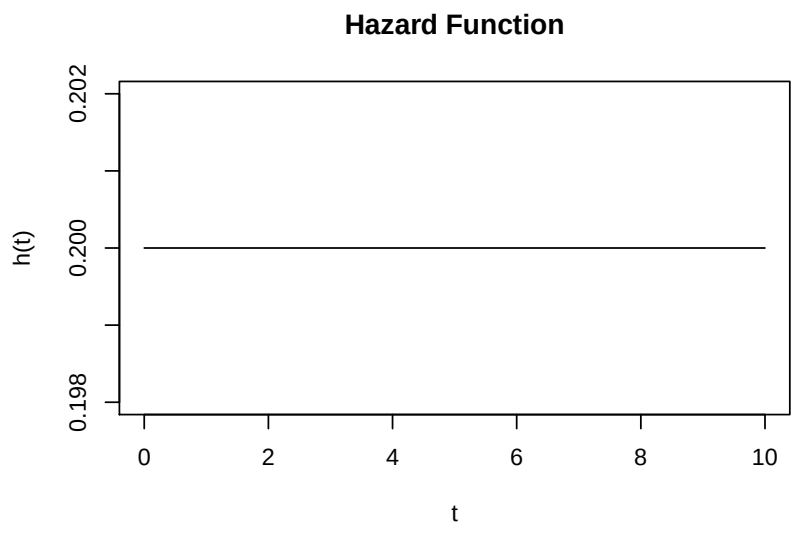
$$h(t) = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}} = \lambda.$$

Exponential Survival Times

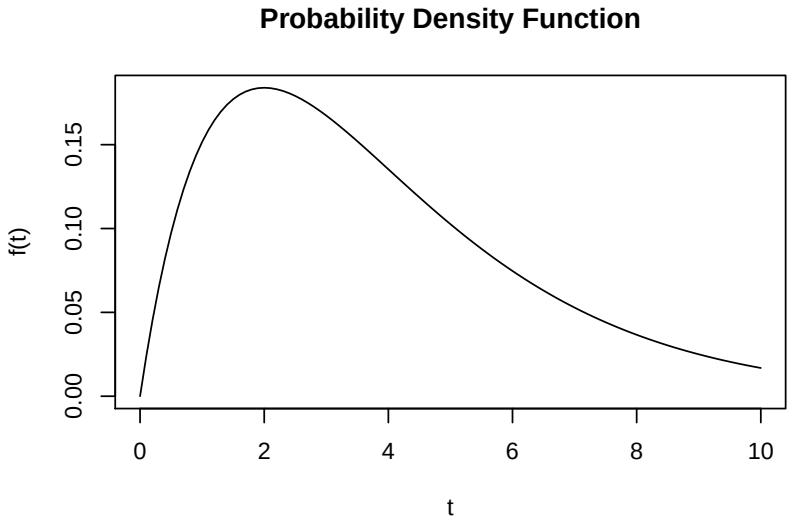


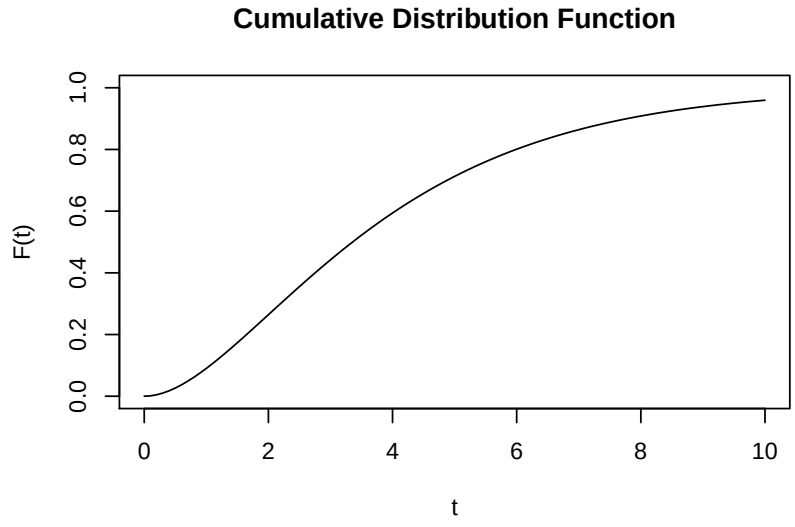


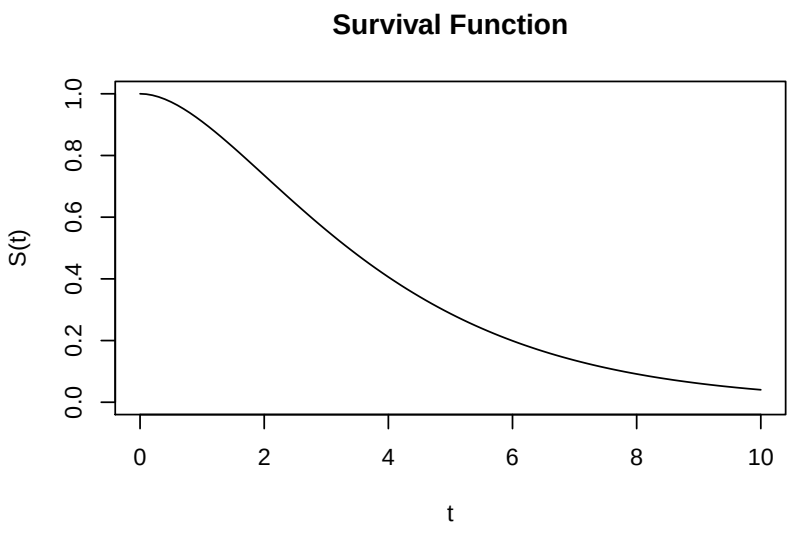


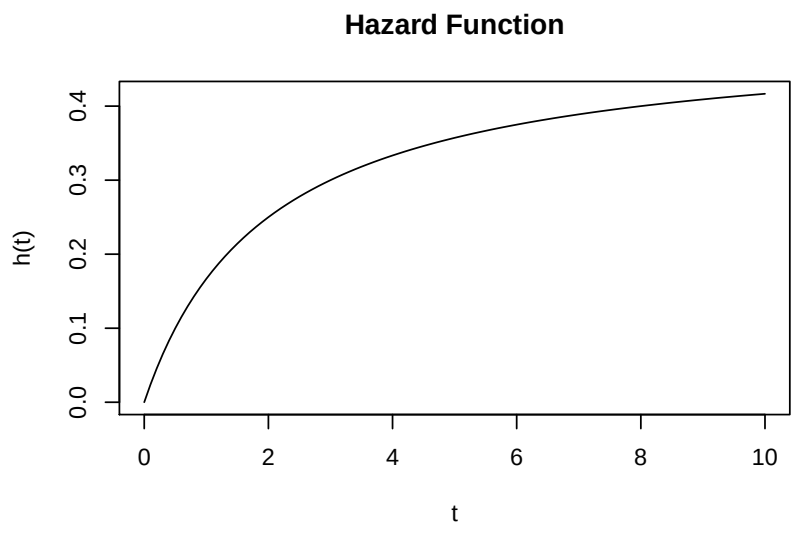


Chi-Squared Survival Times

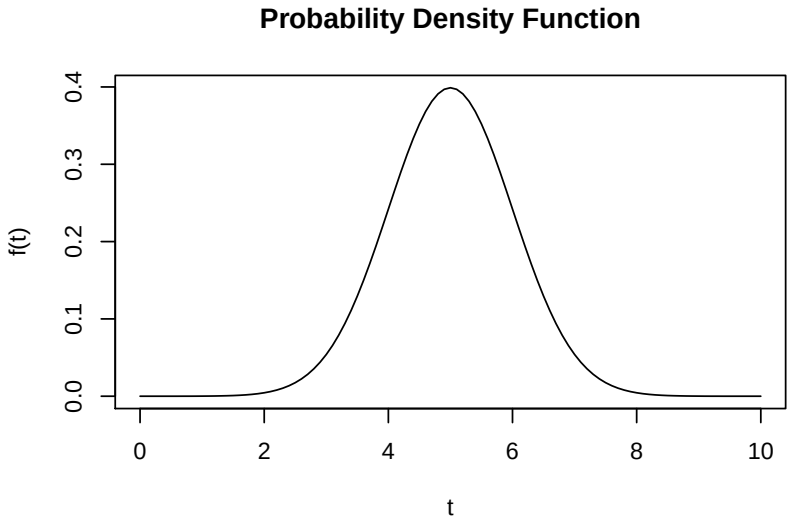


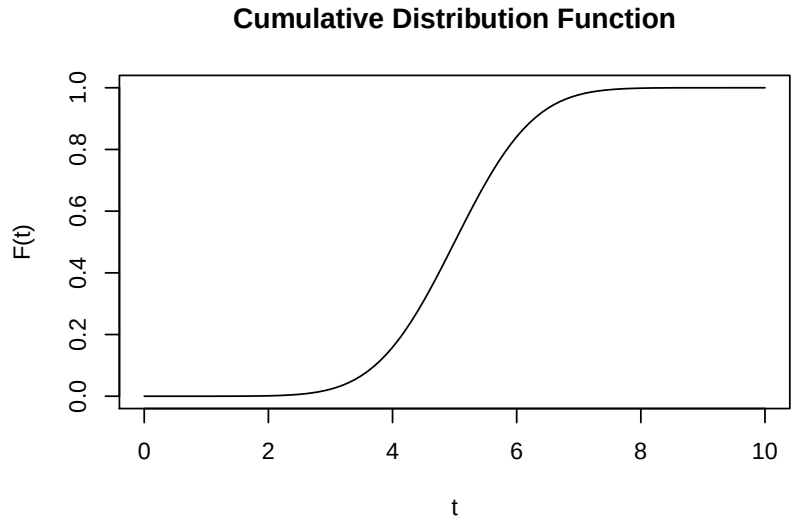


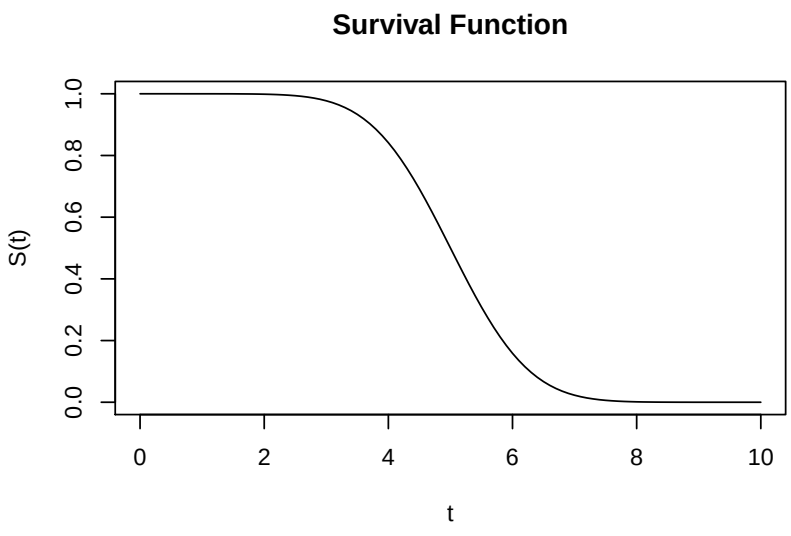


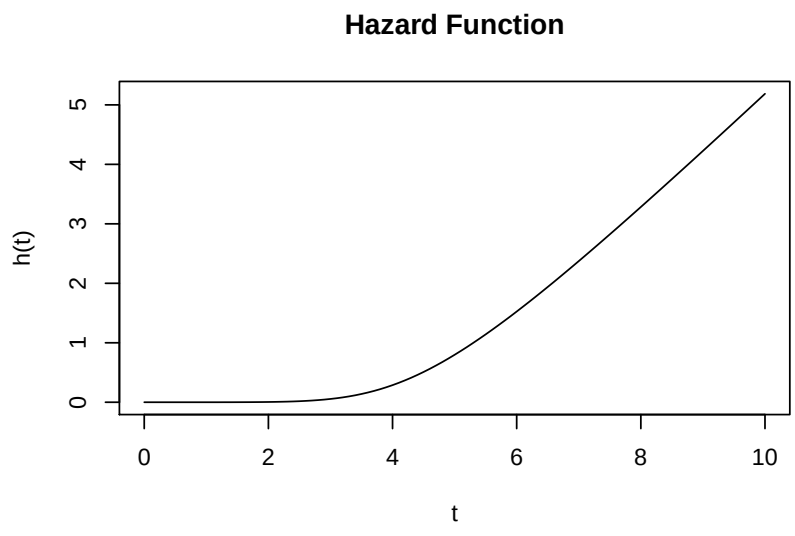


Normal Survival Times

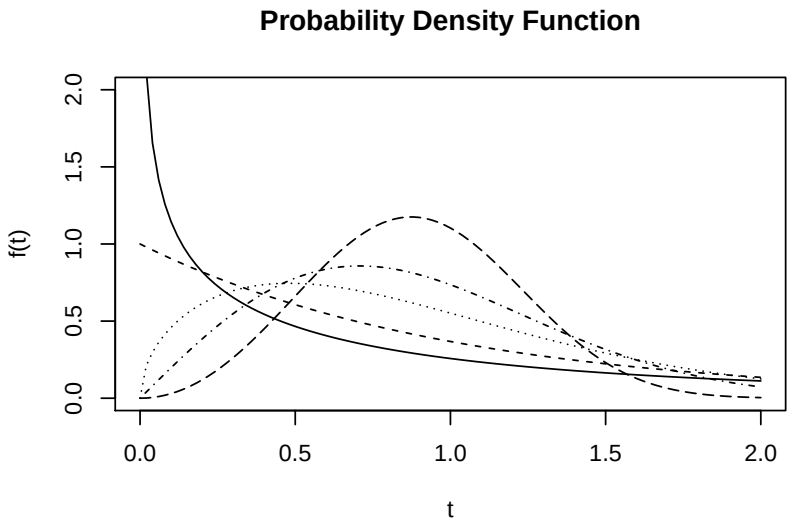


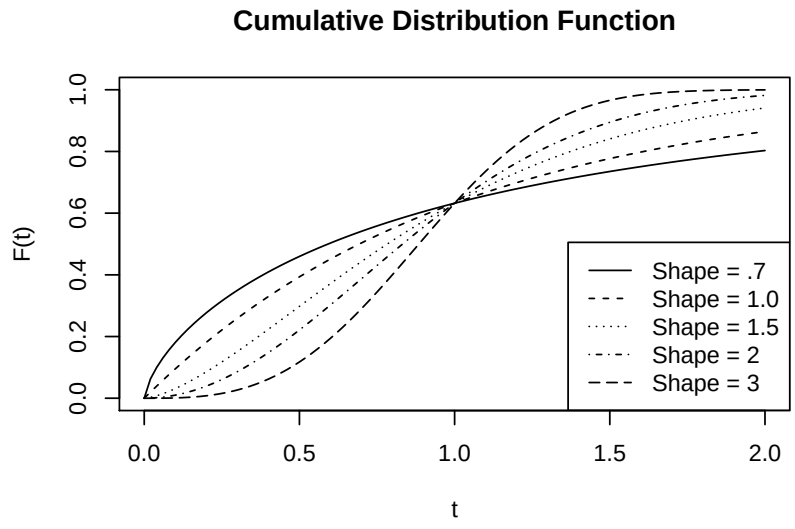


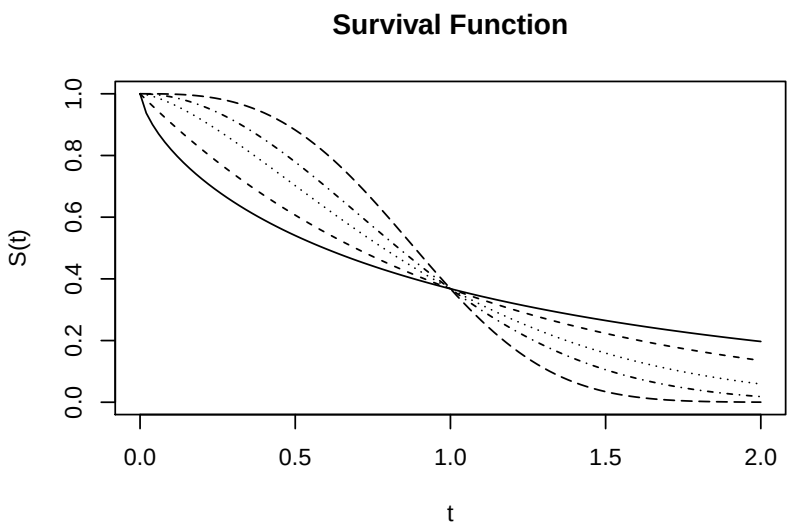


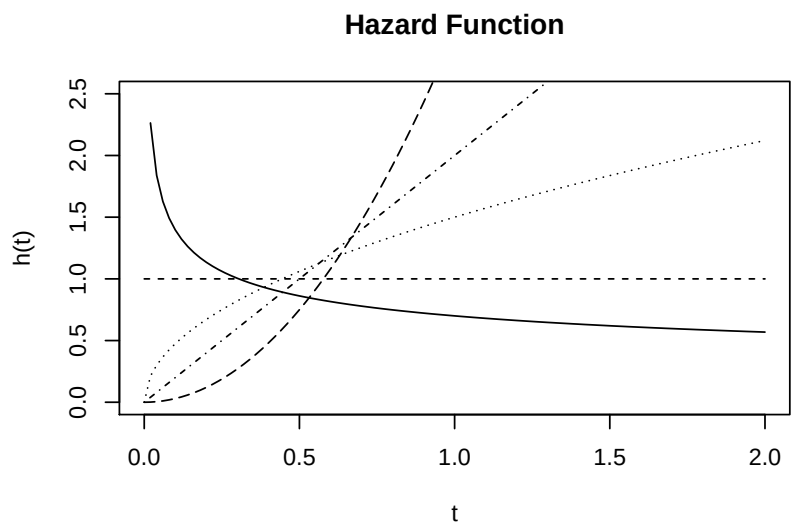


Weibull Survival Times







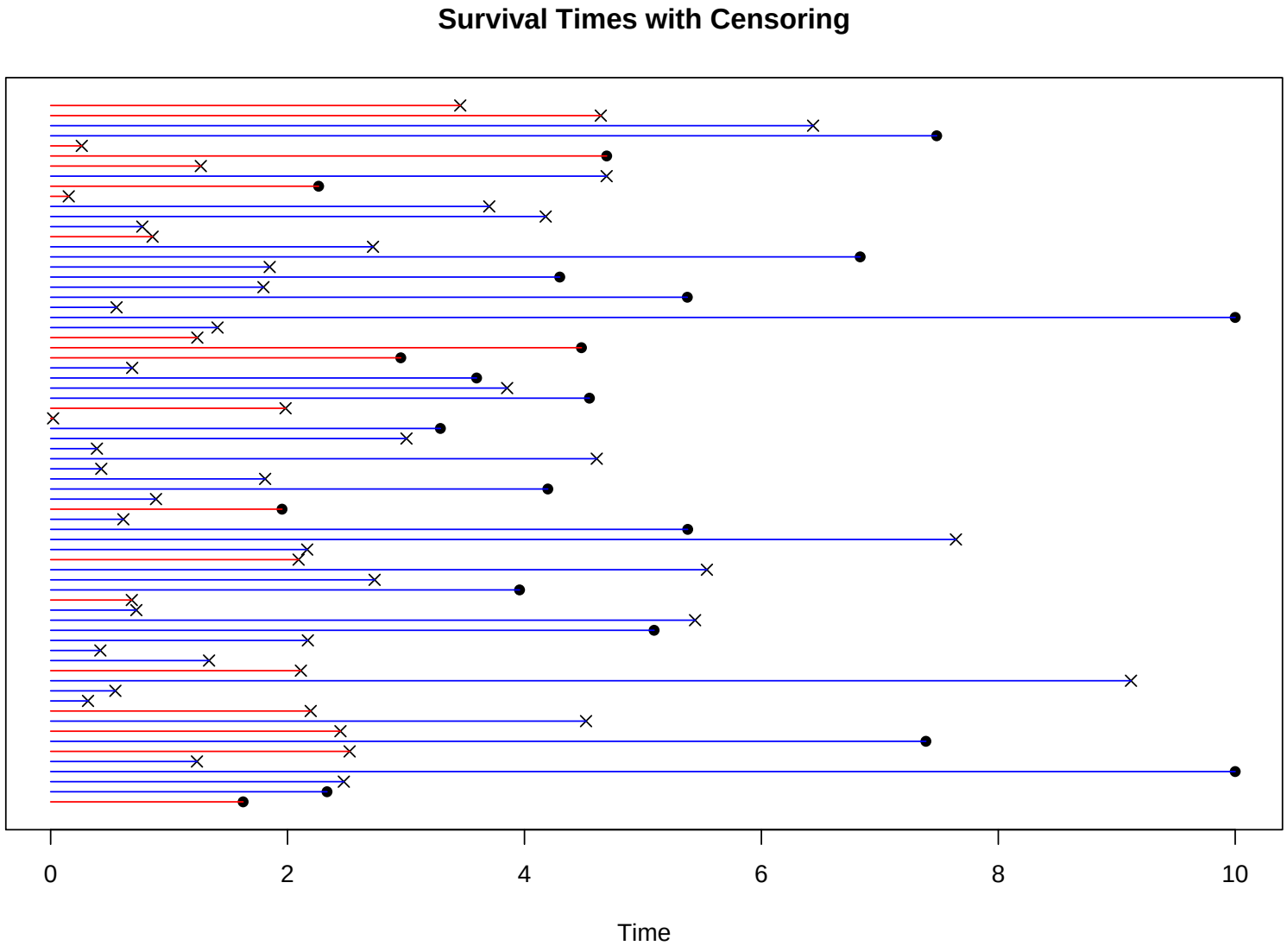


Simulating New Data Using Factors

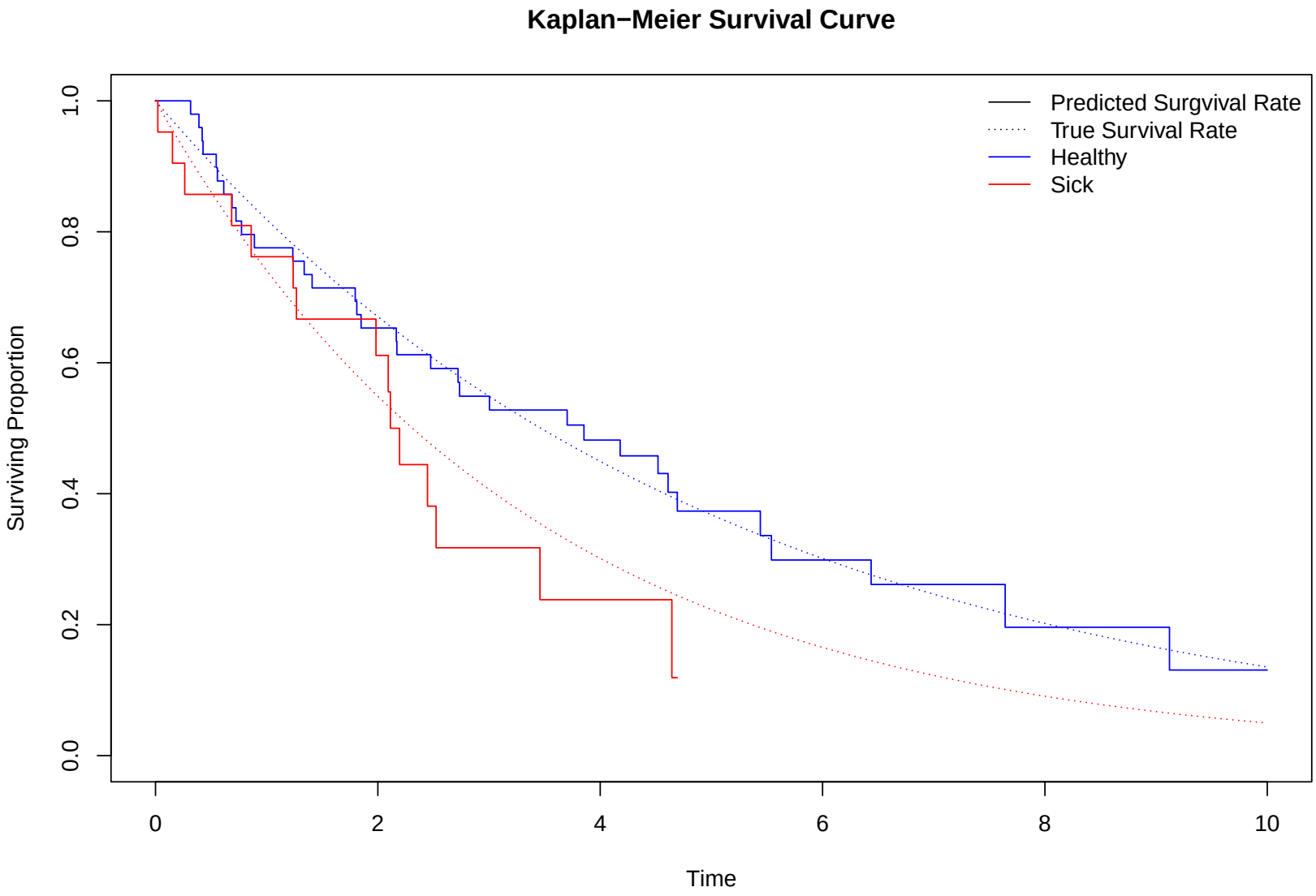
- A larger sample size to get an idea for the curve of the sick individuals.

```
n = 70
lambda.sick = .3
sick = rbinom(n, 1, .3)
lifetime.healthy = rexp(n, rate = lambda)
lifetime.sick = rexp(n, rate = lambda.sick)
lifetime = ifelse(sick, lifetime.sick, lifetime.healthy)
censor = pmin(1 + rexp(n, rate = .1), 10)
endtime = pmin(lifetime, censor)
death = censor > lifetime
data = cbind.data.frame(endtime, death, sick)
data.surv = Surv(data$endtime,data$death)
```

Data Including Sickness Factor



Survival Curves



Log-Rank Test for Difference

H_0 : There is no difference between survival curves for sick individuals.

```
data.test = survdiff(data.surv ~ data$sick)
data.test
```

```
## Call:
## survdiff(formula = data.surv ~ data$sick)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## data$sick=0 49         34     38.8     0.586       2.98
## data$sick=1 21         15     10.2     2.220       2.98
##
##  Chisq= 3  on 1 degrees of freedom, p= 0.08
```

Regression

The hazard function is proportional risk, and can never be negative. An appealing idea is to perform a regression on the hazard function

$$h_i(t) = h_0(t)e^{\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}}.$$

If there is a good reason to suspect a specific baseline hazard function $h_0(t)$, this model is reasonable and a paremetric regression makes sense. Note, the Weibull distribution can provide many different shapes of hazard function.

Cox Regression

Also calleed the Proportional Hazards Model, Cox Regression allows for covariates to be considered. We again want to model the log of the hazard function

$$\log h_i(t) = \log h_0(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}.$$

Unfortunately, we would need to specify the right baseline hazard function $h_0(t)$ to get results, so instead we look at ratios

$$\begin{aligned} HR_{ij} &= \frac{h_i(t)}{h_j(t)} \\ &= \frac{h_0(t)\exp(x_i\beta)}{h_0(t)\exp(x_j\beta)} \\ &= e^{(x_i - x_j)\beta}. \end{aligned}$$

References

- Sainani, Kristin Ph.D. *Introduction to Survival Analysis.* Stanford University.
- Sullivan, Lisa Ph.D. *Survival Analysis.* Boston University School of Public Health. June 3, 2016
- Zhou, Mai *Using Software R to do Survival Analysis and Simulation. A tutorial.* Department of Statistics, University of Kentucky.
- *Survival Analysis.* STHDA. 13 Dec 2016.
- Wikipedia contributors. *Survival analysis.* Wikipedia, The Free Encyclopedia, 13 May. 2021.
- Wikipedia contributors. *Proportional hazards model.* Wikipedia, The Free Encyclopedia, 30 Dec. 2020.
- Wikipedia contributors. *Kaplan–Meier estimator.* Wikipedia, The Free Encyclopedia, 10 May. 2021.
- Hosmer, David W., and Lemeshow, Stanley *Survival Analysis: Applications to Opthalmic Research.* Department of Public Health, University of Massachusetts and the College of Public Health, The Ohio State University, Jul 23, 2008.