Rotterdam

JACK TAN, YIMING MIAO

2021-03-08

Contents

Chapter 1

Motivation

According to World Health Organization∗, Cancer is a leading cause of death worldwide, accounting for an estimated 9.6 million deaths in 2018. And amongst all cancer types, breast cancer(along with lung cancer) has the top cases of death: 2.09 million cases in 2018. According to the CDC∗, Breast cancer is also the second most common cancer among women in the United States, comprising 22.9% of invasive cancers in women and 16% of all female cancers. However, because of the cancer's characteristics, breast cancer patients have relatively high 5-year survival rate of 85% compared to other more lethal cancers according to research conducted in the UK∗. We think it is worthwhile to look at the relationship between survival/recurrence time and some diagnostic criterion. We are also going to explore the effect of different treatments on survival/recurrence.

1.1   Some Background Information

For doctors to be able to assess the severity and different types of breast cancer, researchers have come up with a diagnosing system called the TNM∗ Staging system that is widely used in the diagnostics of breast cancer:

Tumor(T): How large is the primary tumor in the breast?

Node (N): Has the tumor spread to the lymph nodes? If so, where, what size, and how many?

Metastasis (M): Has the cancer spread to other parts of the body?

Generally, the results from the above three features are combined to form a diagnosis of a total of 5 stages of breast cancer: stage 0 (zero), which is non-invasive ductal carcinoma in situ (DCIS), and stages I through IV (1 through 4), which are used for invasive breast cancer.

We will be using data related to this system, especially Tumor and Node(since metastasis is often not recorded in datasets), to conduct our exploration.

Chapter 2

Data Exploration

2.1 Loading Data

```r
data(rotterdam)
```

The data that we are going to use is called `rotterdam`, and it is a dataset that's pre-recorded in the survival package. According to the documentation of the package, the data are retrieved from the Rotterdam tumor bank, which include various anonymous information about patients with breast cancer. Below is a table of the variables in the dataset:

Variable name	Description
pid	patient identifier
year	year of cancer incidence
age	age
meno	menopausal status (0= premenopausal, 1= postmenopausal)
size	tumor size, a factor with levels <=20, 20-50, >50
grade	tumor grade
nodes	number of positive lymph nodes
pgr	progesterone receptors (fmol/l)
er	estrogen receptors (fmol/l)
hormon	hormonal treatment (0=no, 1=yes)
chemo	chemotherapy
rtime	days to recurrence or last follow-up
recur	0= no recurrence, 1= recurrence
dtime	days to death or last follow-up
death	0= alive, 1= dead

From the description above, we see that there are `size` which stands for the size of the tumor, `nodes` which stands for how many lymph nodes are test cancer positive, so we have 2 criterions out of the three suggested in the background info.

2.2   Data Wrangling

```
rotterdam %>%
  group_by(size) %>%
  summarise(number = n(), .groups = 'drop')
```

```
## # A tibble: 3 x 2
##   size  number
##   <fct> <int>
## 1 <=20   1387
## 2 20-50  1291
## 3 >50     304
```

For lymph nodes, the usual way of classifying the severity would be:`N0` for no positive nodes; `N1` for 1-3 positive nodes; `N2` for 4-9 positive nodes; and `N3` for more than 10 nodes. We will follow this classification and make a new factor called `Nodes_level`

```
rotterdam <- rotterdam %>%
  mutate(Nodes_level = ifelse(nodes == 0, "N0",
                        ifelse(nodes >= 1 & nodes <= 3, "N1",
                          ifelse(nodes >= 4 & nodes <= 9, "N2",
                            ifelse(nodes >= 10, "N3", NaN)))))
rotterdam %>%
  group_by(Nodes_level) %>%
  summarise(number = n(), .groups = 'drop')
```

```
## # A tibble: 4 x 2
##   Nodes_level number
##   <chr>        <int>
## 1 N0            1436
## 2 N1             764
## 3 N2             515
## 4 N3             267
```

Since the `grade` variable in our dataset is a numeric variable whereas we actually want to treat it as a factor, we do the following:

```
rotterdam <- rotterdam %>%
  mutate(grade = as.factor(grade))
```

As we were examining through the data, we found that upon the `chemo` variable and the `hormon` variable, there are instances where patients gets both therapy or neither. So in order to explore the relationship between treatment and survival, we introduce a new variable called `Treatment`, using the `chemo` and `hormon` variables.
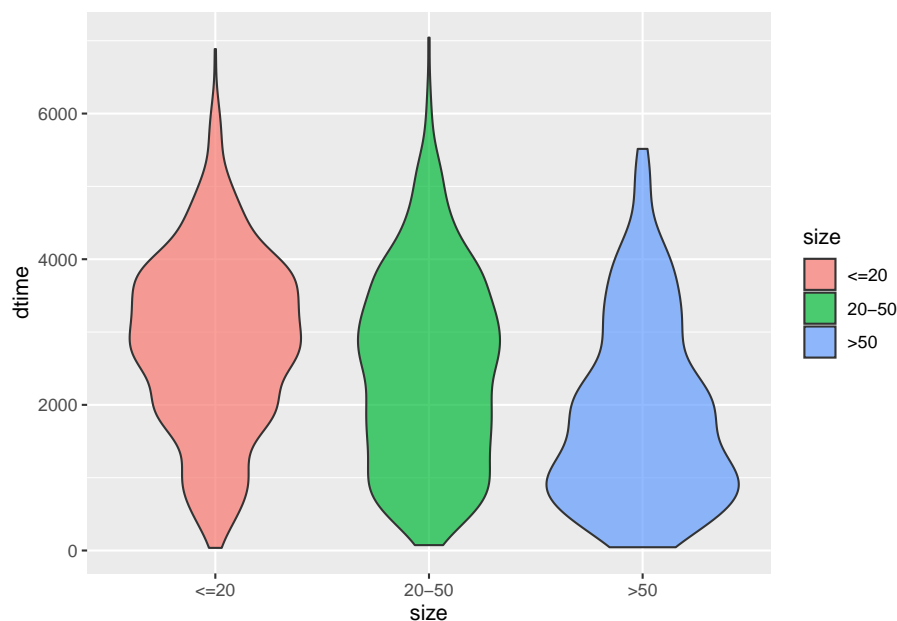
```
rotterdam <- rotterdam %>%
  mutate(Treatment = ifelse(chemo == 1 & hormon == 0, "Chemo", ifelse(chemo == 0 & hormon == 1, '
```

2.3 Data visualizations

2.3.1 Diagnositcs vs. Survival Times

`size` vs. `dtime`

```
ggplot(data = rotterdam, aes(x = size, y = dtime, fill = size)) +
  geom_violin(alpha = 0.7)
```
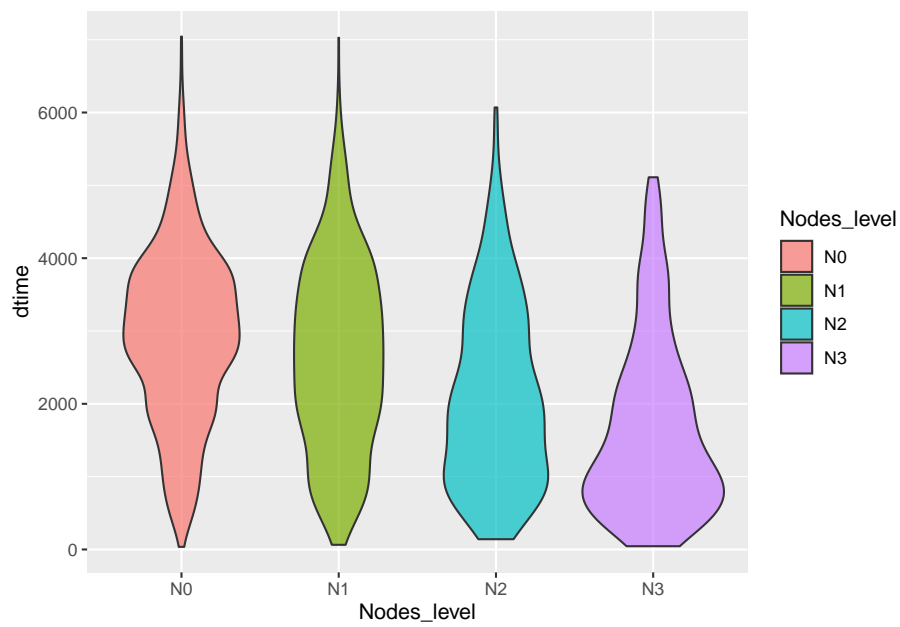

`size` vs. `rtime`

```r
ggplot(data = rotterdam, aes(x = size, y = rtime, fill = size)) +
  geom_violin(alpha = 0.7)
```
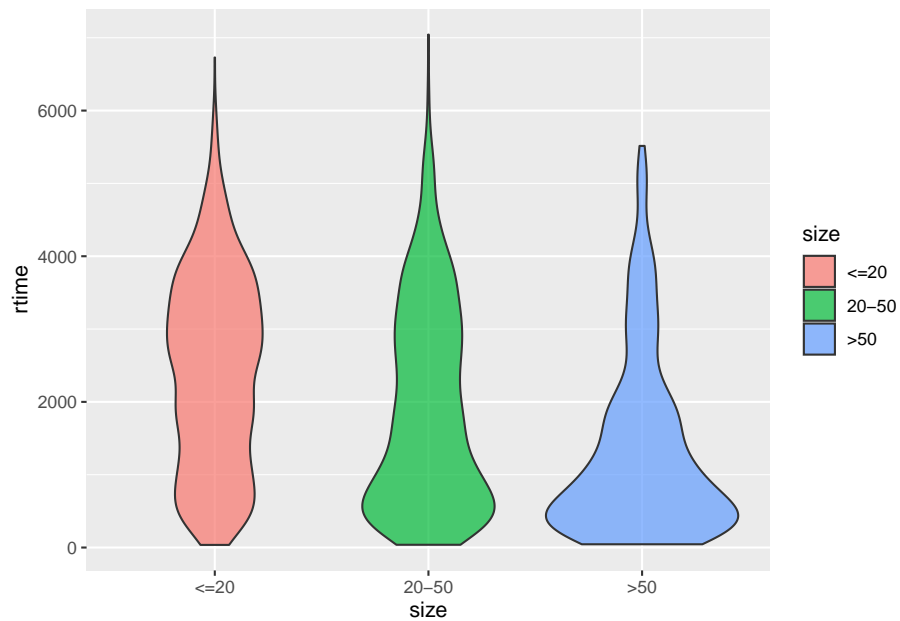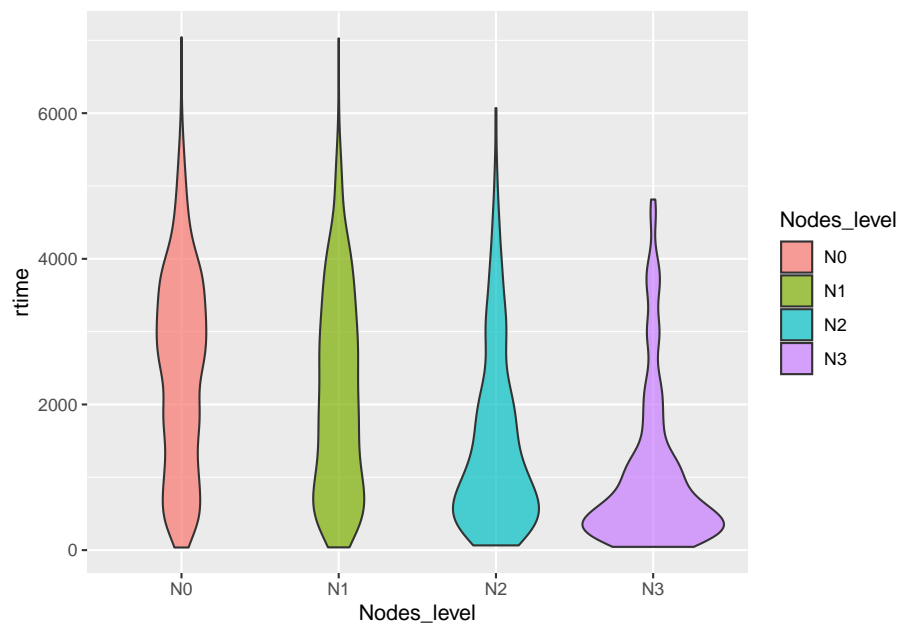

Nodes_level vs. dtime

```r
ggplot(data = rotterdam, aes(x = Nodes_level, y = dtime, fill = Nodes_level)) +
  geom_violin(alpha = 0.7)
```
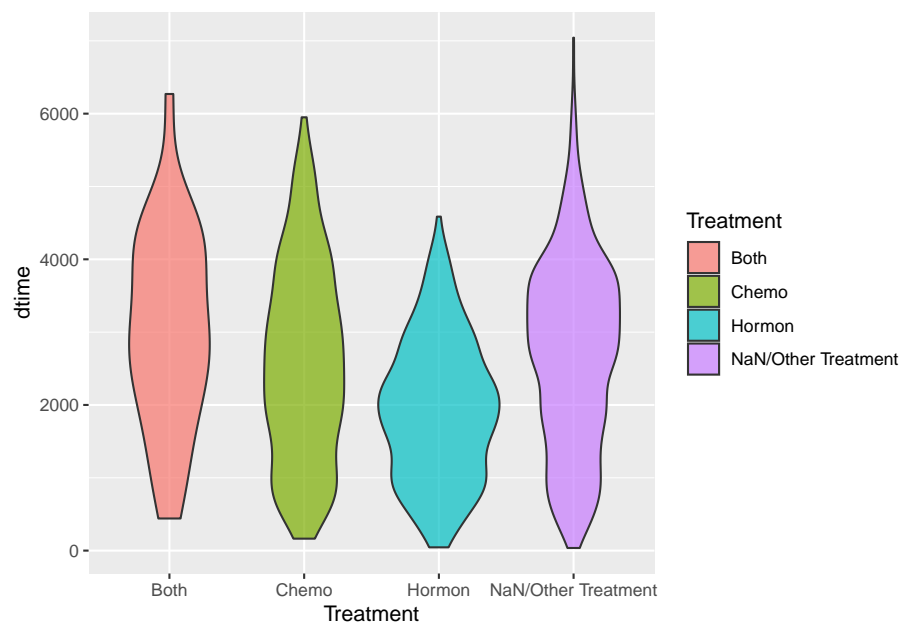
Nodes_level vs. `rtime`

```r
ggplot(data = rotterdam, aes(x = Nodes_level, y = rtime, fill = Nodes_level)) +
  geom_violin(alpha = 0.7)
```
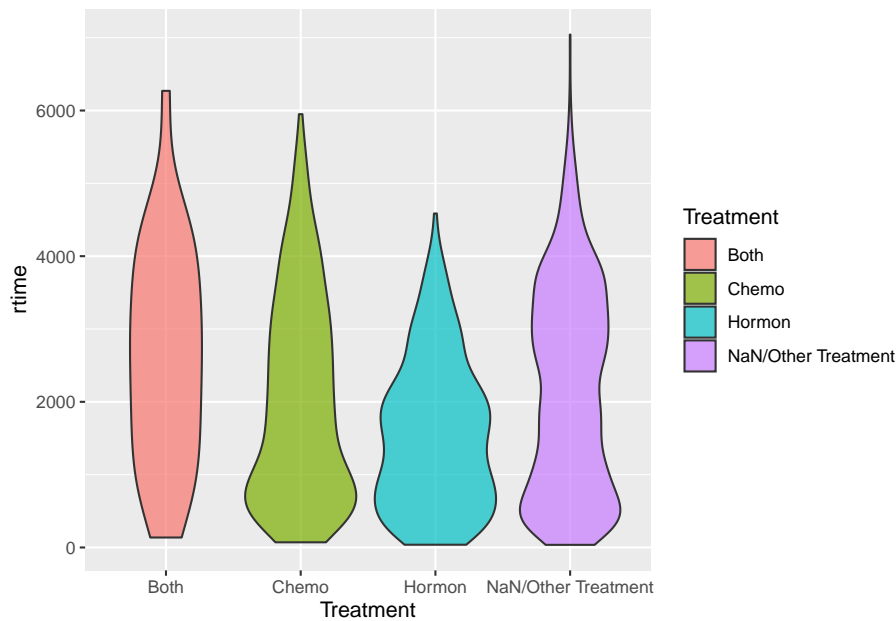
2.3.2   Treatment vs. Survival Times

Treatment vs. dtime

```
ggplot(data = rotterdam, aes(x = Treatment, y = dtime, fill = Treatment)) +
  geom_violin(alpha = 0.7)
```


Treatment vs. rtime

```
ggplot(data = rotterdam, aes(x = Treatment, y = rtime, fill = Treatment)) +
  geom_violin(alpha = 0.7)
```

A very important criterion in analysis about cancer is the 5-year survival rate. In order to examine that, we introduce a new variable called `5_year_survival`, which indicates 1 if a patients survival time is larger than 5 years and 0 vice versa.

```
rotterdam <- rotterdam %>%
  mutate(dtime_Years = floor(dtime/365)) %>%
  mutate(‘5_year_survival‘ = ifelse(dtime_Years >= 5, 1, 0))
```

Now we want to calculate the 5-year survival rate for the population in the dataset.

```
rotterdam %>%
  group_by(‘5_year_survival‘) %>%
  summarise(number = n(), .groups = 'drop')
```

```
## # A tibble: 2 x 2
##   ‘5_year_survival‘ number
##               <dbl>  <int>
## 1                 0    898
## 2                 1   2084
```

```r
2084/(898+2084)
```

```
## [1] 0.6988598
```

And also the important 10-year survival rate.

```r
rotterdam <- rotterdam %>%
  mutate(dtime_Years = floor(dtime/365)) %>%
  mutate(`10_year_survival` = ifelse(dtime_Years >= 10, 1, 0))
```

Now we calculate the 10-year survival rate for the population in the dataset.

```r
rotterdam %>%
  group_by(`10_year_survival`) %>%
  summarise(number = n(), .groups = 'drop')
```

```
## # A tibble: 2 x 2
##   `10_year_survival` number
##                <dbl>  <int>
## 1                  0   2297
## 2                  1    685
```

```r
685/(685 + 2297)
```
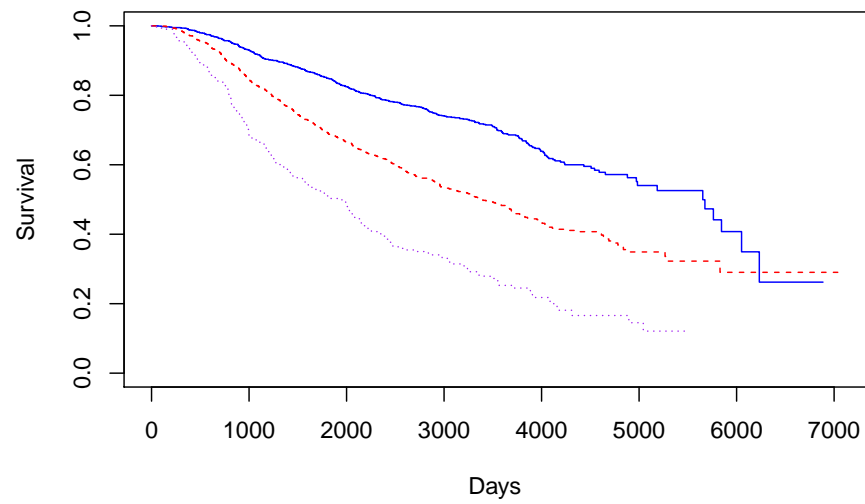
```
## [1] 0.2297116
```

Chapter 3

Survival

3.1 Loading Data

```r
data(rotterdam)
```
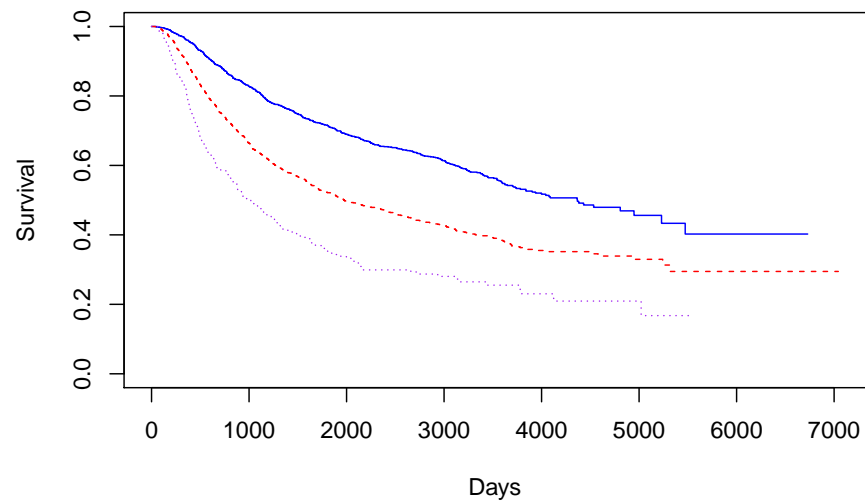
3.2 Kaplan-Miere

3.2.1 `size` vs. Survival Times

size vs. dtime

```r
KM_None_Death <- survfit(Surv(dtime, death) ~ size, data = rotterdam)
plot(KM_None_Death, conf.type = "plain", col = c("blue","red","purple"), xlab="Days", ylab="Survi
```
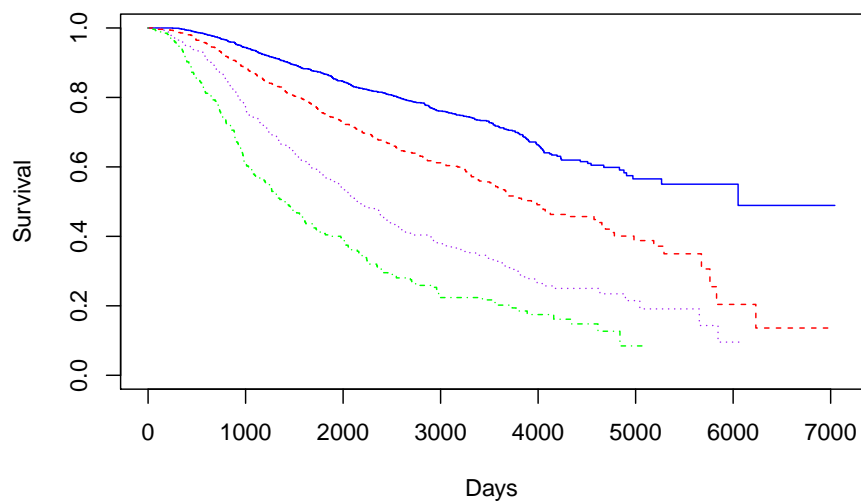
13

size vs. rtime

```
KM_None_Recur <- survfit(Surv(rtime, recur) ~ size, data = rotterdam)
plot(KM_None_Recur, conf.type = "plain", col = c("blue","red","purple"), xlab="Days", y
```
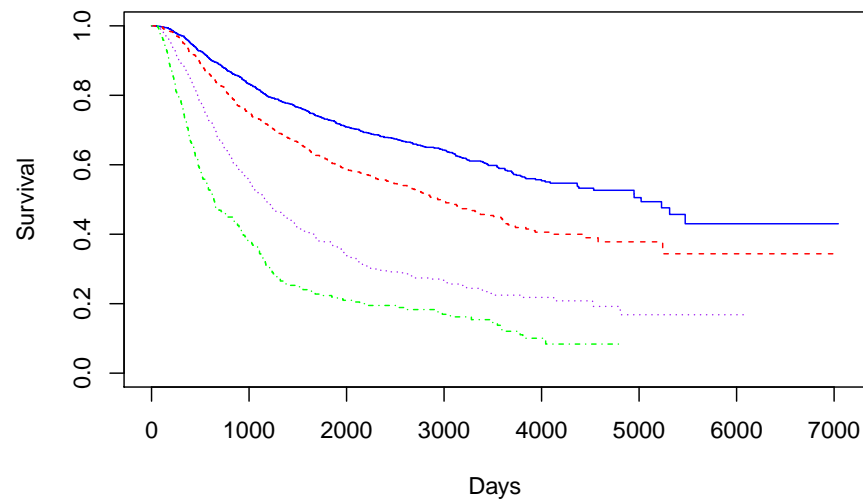
3.2.2 `Nodes_level` vs. Survival Times

size vs. `dtime`

```
KM_None_Death <- survfit(Surv(dtime, death) ~ Nodes_level, data = rotterdam)
plot(KM_None_Death, conf.type = "plain", col = c("blue","red","purple","green"), xlab="Days", yl
```


size vs. `rtime`

```
KM_None_Recur <- survfit(Surv(rtime, recur) ~ Nodes_level, data = rotterdam)
plot(KM_None_Recur, conf.type = "plain", col = c("blue","red","purple", "green"), xlab="Days", yl
```

3.2.3   Treatment vs. Survival Times

Treatment vs. dtime

```
KM_Treatment_Death <- survfit(Surv(dtime, death) ~ Treatment, data = rotterdam)
plot(KM_Treatment_Death, conf.int = FALSE, col = c("blue", "red", "purple", "green"),
legend(1, 0.4, legend=c("Both", "Chemo","Hormon", "NaN/Other Treatment"),
       col=c("blue", "red", "purple", "green"), lty=1:4, cex=0.8,
       title="Treatment Group", text.font=6)
```
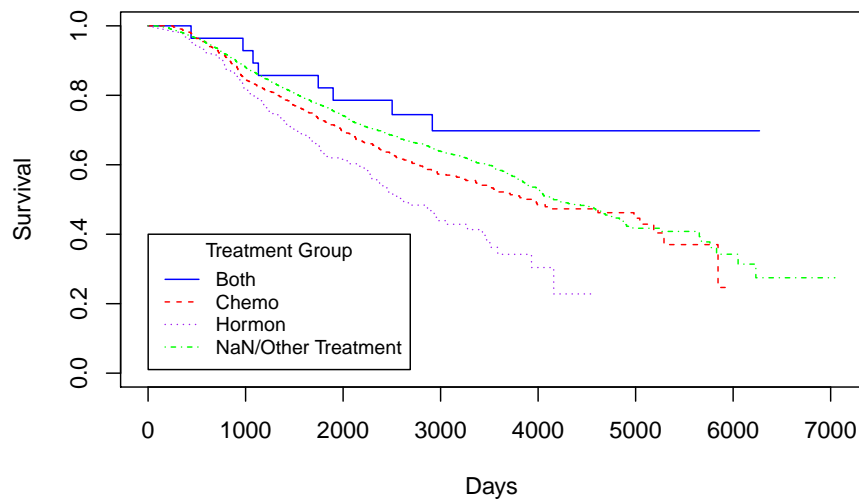
Treatment vs. rtime

```
KM_Treatment_Recur <- survfit(Surv(rtime, recur) ~ Treatment, data = rotterdam)
plot(KM_Treatment_Recur, conf.int = FALSE, col = c("blue", "red", "purple", "green"), xlab="Days"
legend(1, 0.4, legend=c("Both", "Chemo","Hormon", "NaN/Other Treatment"),
       col=c("blue", "red", "purple", "green"), lty=1:4, cex=0.8,
       title="Treatment Group", text.font=6)
```
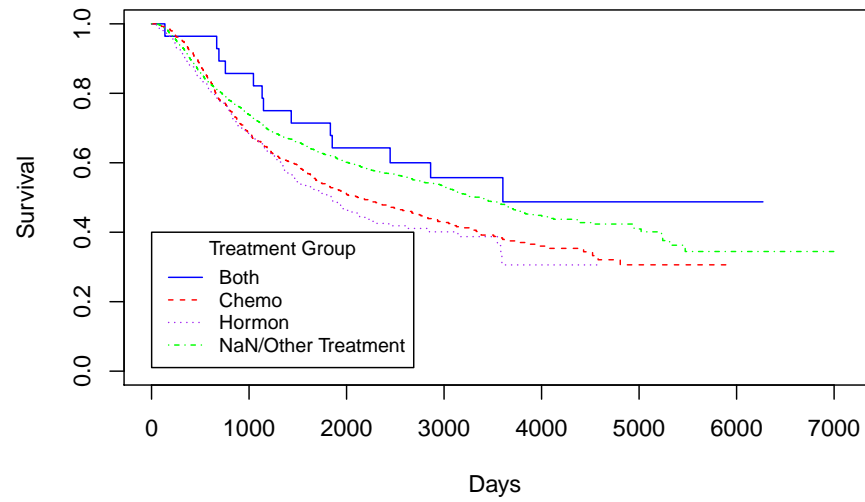
3.3   Parametric Models

3.4   Cox-PH

Chapter 4

Reference

- "Breast Cancer - Stages." Cancer.Net, 14 Aug. 2020, www.cancer.net/cancer-types/breast-cancer/stages#:~:text=There are 5 stages of,to plan the best treatments.

- "Breast Cancer Statistics." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 8 June 2020, www.cdc.gov/cancer/breast/statistics/index.htm.

- "Breast Cancer Survival Statistics." Cancer Research UK, 22 Jan. 2021, www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/survival#heading-Zero.

- "Cancer." World Health Organization, World Health Organization, www.who.int/news-room/fact-sheets/detail/cancer.