Rotterdam

JACK TAN, YIMING MIAO

2021-03-09

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. Finally we will look at all the factors together.

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

```
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.

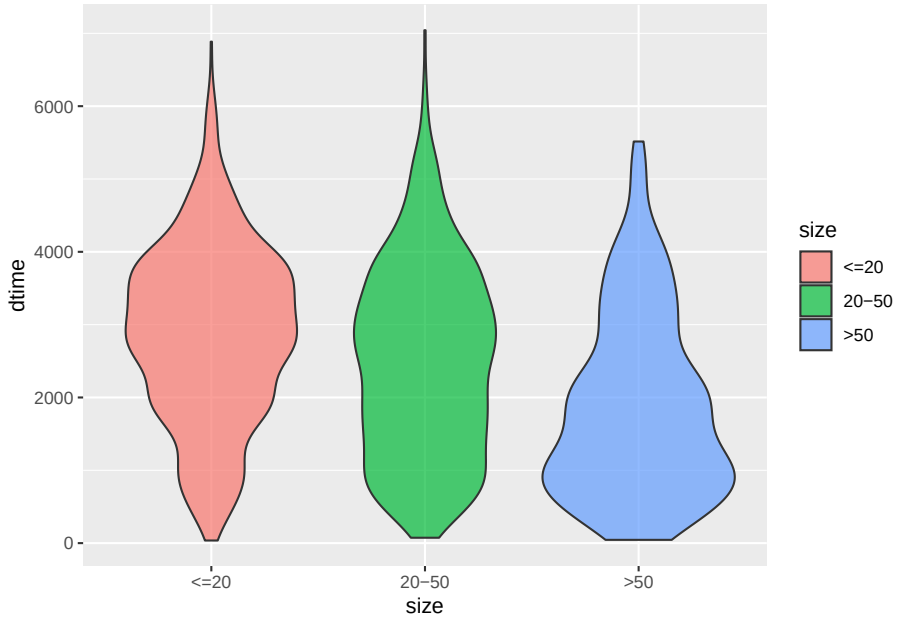
```
rotterdam <- rotterdam %>%
  mutate(Treatment = ifelse(chemo == 1 & hormon == 0, "Chemo",
                     ifelse(chemo == 0 & hormon == 1, "Hormon",
                     ifelse(chemo == 1 & hormon == 1, "Both", "NaN/Other Treatment")))) %>%
  mutate(Treatment = as.factor(Treatment))
```

2.3 Data visualizations and exploration

2.3.1 Diagnostics vs. Survival Times

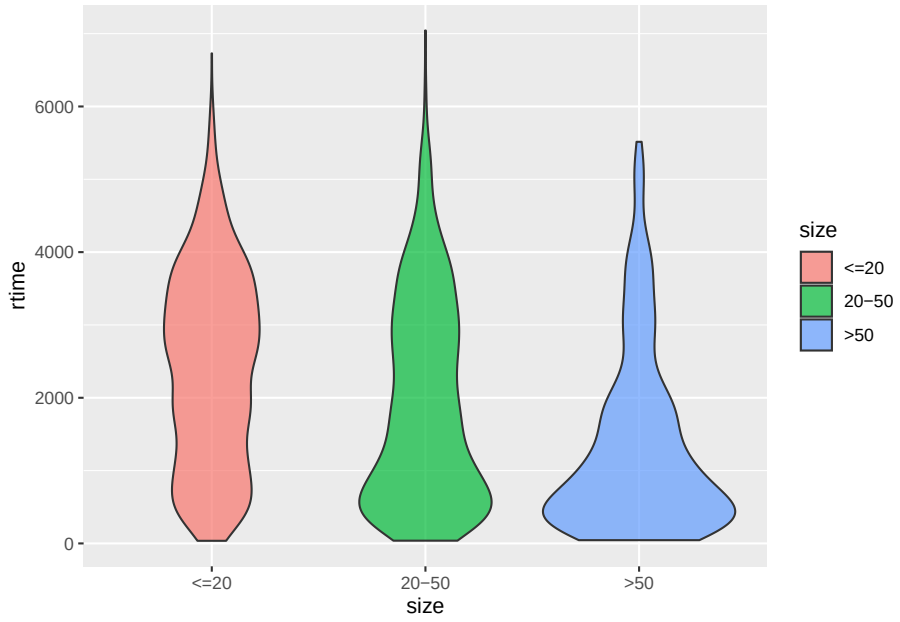
`size` vs. `dtime`

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



`size` vs. `rtime`

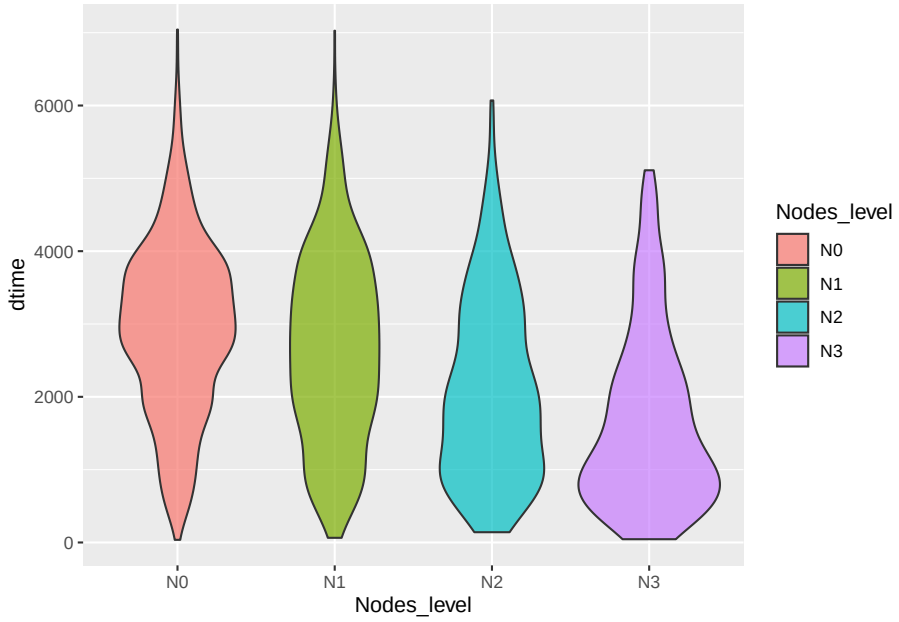
```
ggplot(data = rotterdam, aes(x = size, y = rtime, fill = size)) +
  geom_violin(alpha = 0.7)
```



As we can see from the two plots above, tumor size could be an important factor that affects patients' survival time and recur time. For `size` smaller than 20, most of the patients are able to survive or encounter recurrence after roughly 3000 days. But for `size` 20-50 and >50, it's highly likely for breast cancer to recur in 500 days.

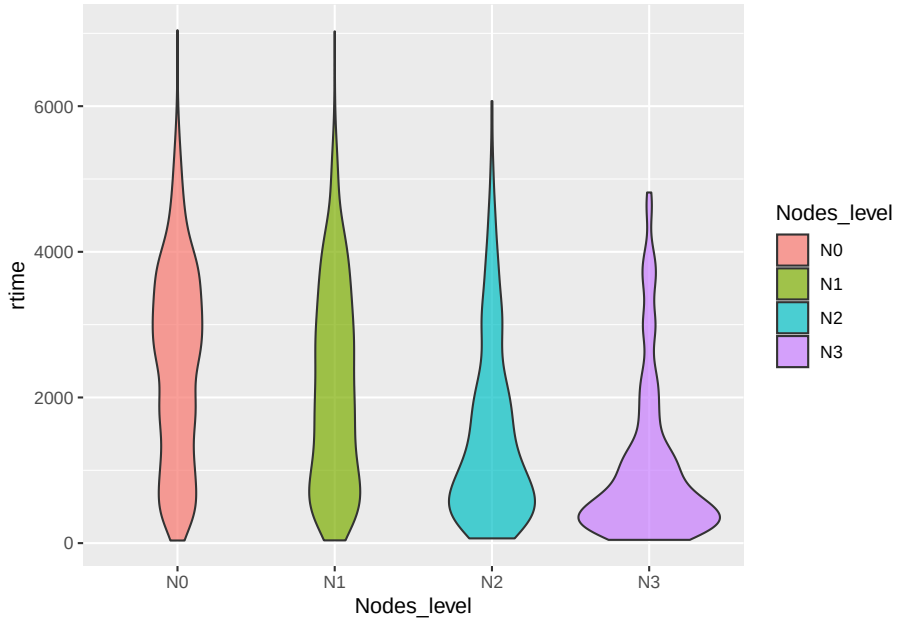
`Nodes_level` vs. `dtime`

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



`Nodes_level` vs. `rtime`

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

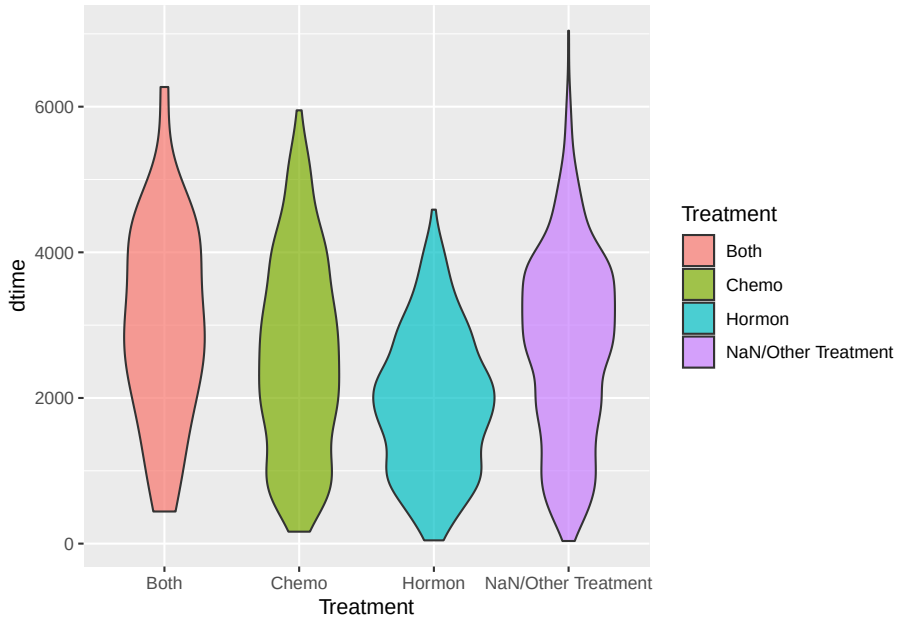


Similarly, `nodes` is also a factor impacting the life of breast cancer patients. For patients with `Nodes_level` N2 and N3, a quite large proportion of them fail to survival more than 1000 days and encounter recurrence in less than 500 days.

2.3.2 Treatment vs. Survival Times

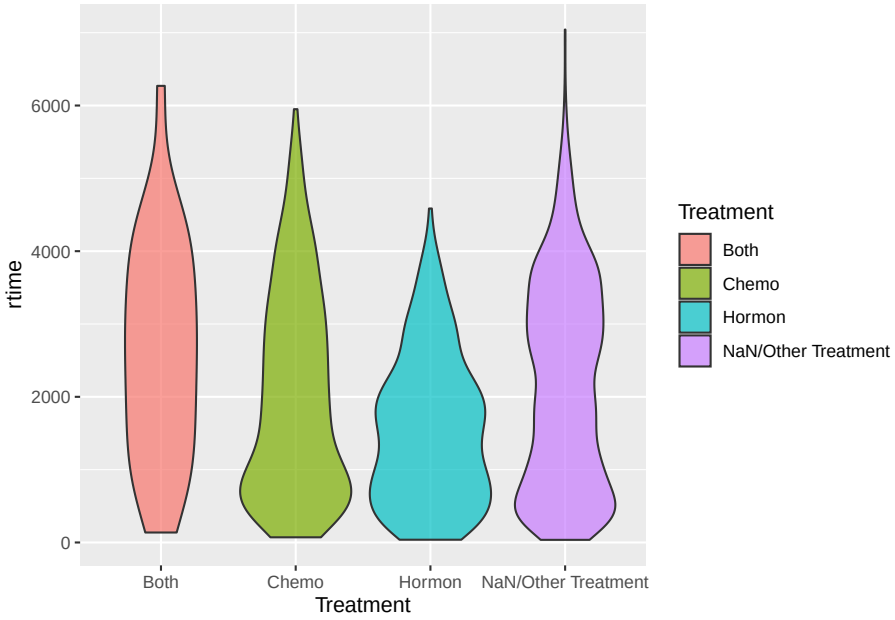
`Treatment` vs. `dtime`

```
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)
```



By examing the two plots above, it seems that `Treatment` would not affect patients survival time or recurrence that much. We can find that `chemo + hormon` is likely to be the one with best curative effect, that patients receiving both chemo and hormon therapy tend to have longer survival time and longer time to recurrence. And the effect of hormon therapy itself seems not that satisfying.

We also found a bi-model shape in `NaN/Other Traetment` group. This may because the two peaks correponds to no treatment and other treatment separately, but currently we don't more information investigating the true reason.

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
```

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

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

And also the important 10-year survival rate.

```
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.

```
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
```

```
685/(685 + 2297)
```

```
## [1] 0.2297116
```

We can find that the 5-year survival rate for breast cancer is just fine, and around 70% of patients are able to live more than 5 years. However, the 10-year survival rate is still disappointing given the current medical level, and only around 20% patients could live more than 10 years after diagnosis.

All above information reminds people, especially females, that screening does save lives. Get screening and diagnosis earlier could greatly help improving the quality of life in the future.

Chapter 3

Survival

3.1 Loading Data

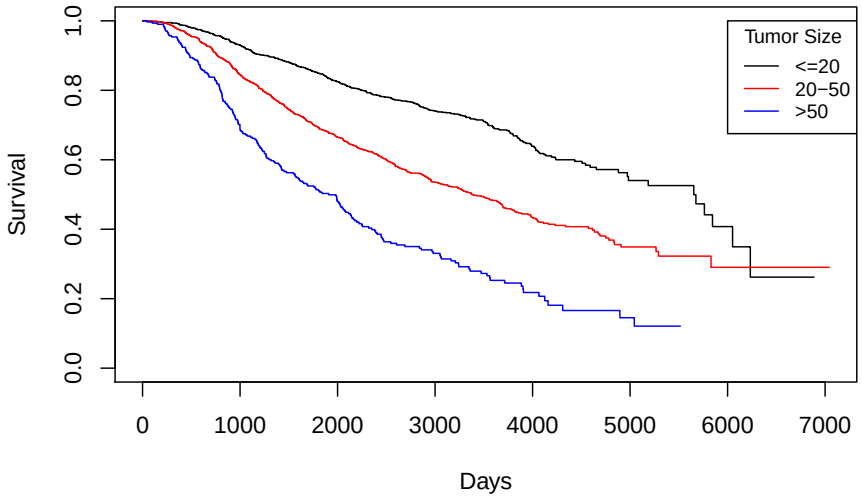
```
data(rotterdam)
```

3.2 Kaplan-Miere

3.2.1 `size` vs. Survival Times

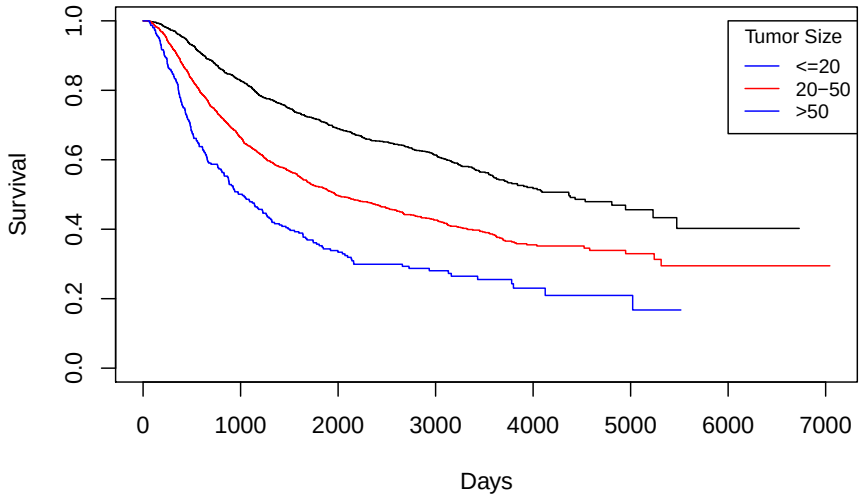
`size` vs. `dtime`

```
KM_None_Death <- survfit(Surv(dtime, death) ~ size, data = rotterdam)
plot(KM_None_Death, conf.type = "plain", col = c("black","red","blue"), xlab="Days", ylab="Surviv
legend(6000, 1, legend=c("<=20", "20-50", ">50"),
       col=c("black", "red", "blue"), lty=1, cex=0.8,
       title="Tumor Size", text.font=6)
```



`size` vs. `rtime`

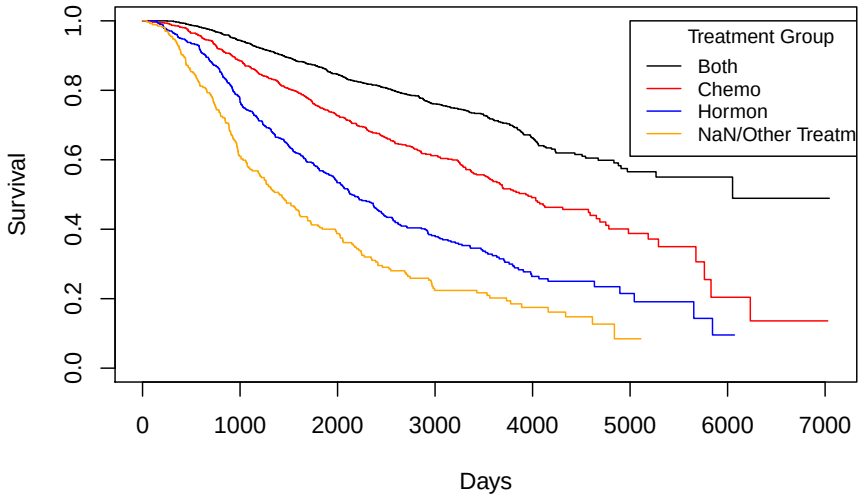
```
KM_None_Recur <- survfit(Surv(rtime, recur) ~ size, data = rotterdam)
plot(KM_None_Recur, conf.type = "plain", col = c("black","red","blue"), xlab="Days", yl
legend(6000, 1, legend=c("<=20", "20-50", ">50"),
       col=c("blue", "red", "blue"), lty=1, cex=0.8,
       title="Tumor Size", text.font=6)
```



3.2.2 Nodes_level vs. Survival Times

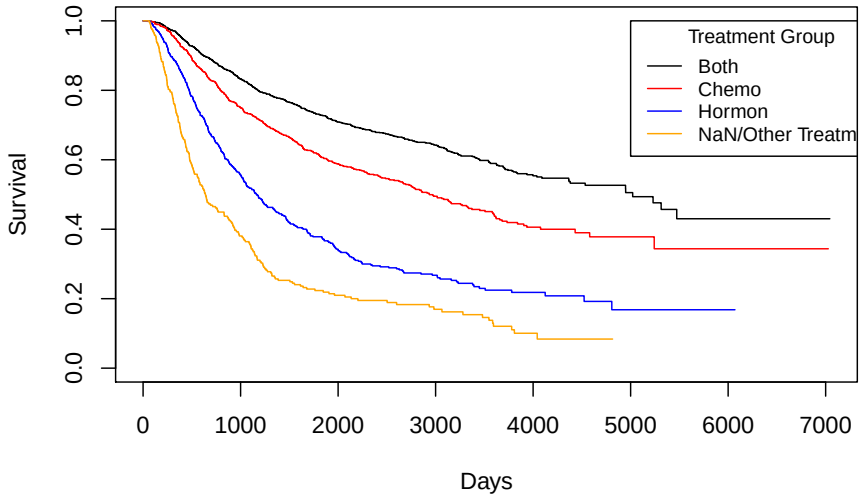
Nodes_level vs. dtime

```
KM_None_Death <- survfit(Surv(dtime, death) ~ Nodes_level, data = rotterdam)
plot(KM_None_Death, conf.type = "plain", col = c("black","red","blue","orange"), xlab="Days", yla
legend(5000, 1, legend=c("Both", "Chemo","Hormon", "NaN/Other Treatment"),
       col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,
       title="Treatment Group", text.font=6)
```



`Nodes_level` vs. `rtime`

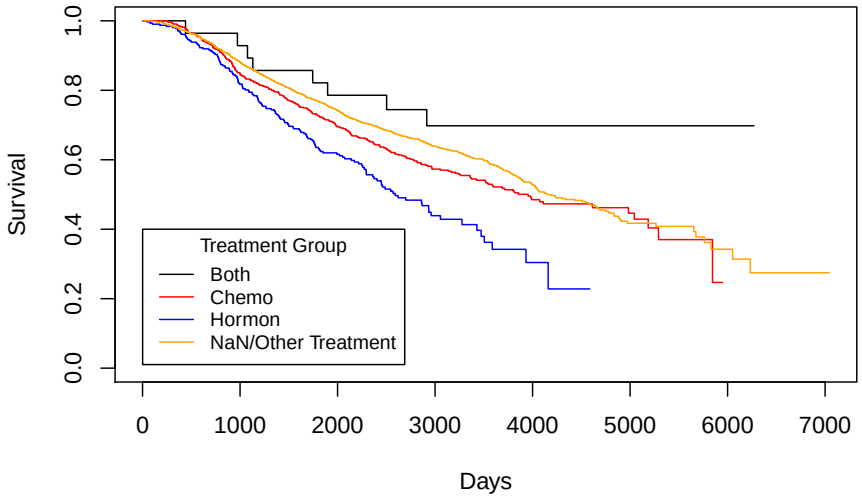
```
KM_None_Recur <- survfit(Surv(rtime, recur) ~ Nodes_level, data = rotterdam)
plot(KM_None_Recur, conf.type = "plain", col = c("black","red","blue","orange"), xlab="
legend(5000, 1, legend=c("Both", "Chemo","Hormon", "NaN/Other Treatment"),
       col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,
       title="Treatment Group", text.font=6)
```



3.2.3 Treatment vs. Survival Times

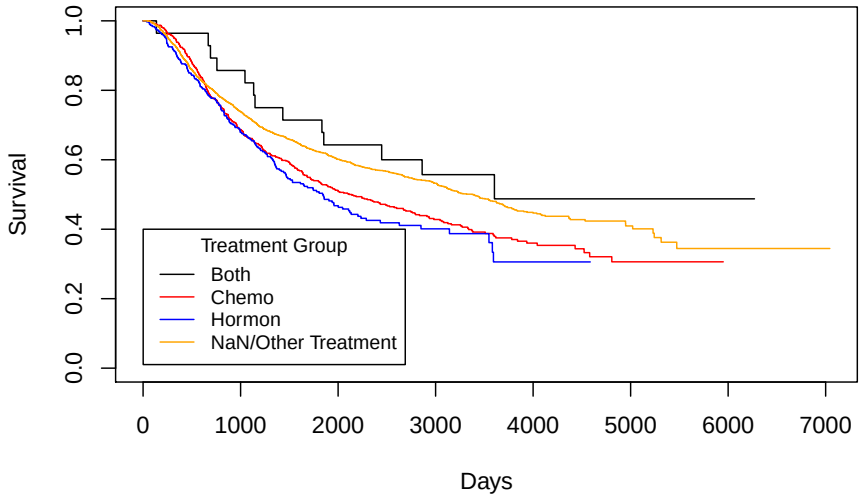
Treatment vs. dtime

```
KM_Treatment_Death <- survfit(Surv(dtime, death) ~ Treatment, data = rotterdam)
plot(KM_Treatment_Death, conf.int = FALSE, col = c("black", "red", "blue", "orange"), xlab="Days"
legend(1, 0.4, legend=c("Both", "Chemo","Hormon", "NaN/Other Treatment"),
       col=c("black", "red", "blue", "orange"), lty=1, 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("black", "red", "blue", "orange"), x
legend(1, 0.4, legend=c("Both", "Chemo","Hormon", "NaN/Other Treatment"),
       col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,
       title="Treatment Group", text.font=6)
```



3.3 Parametric Models

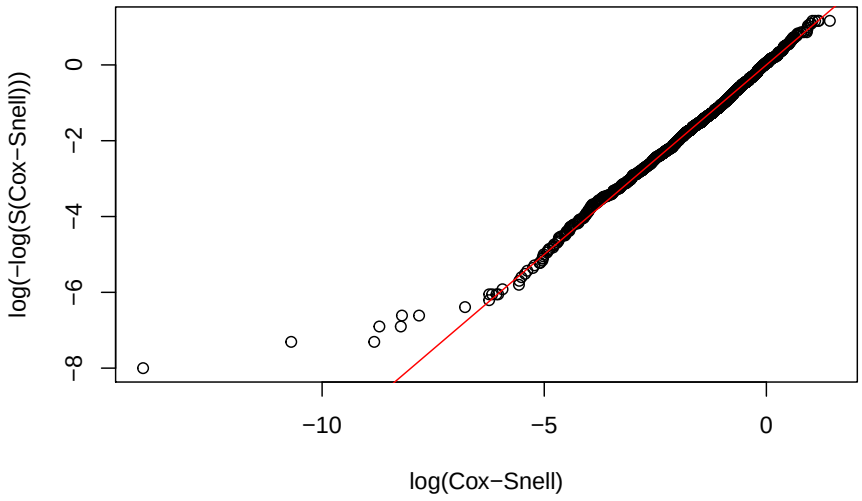
```
# The Cox-Snell function takes as inputs
# 1. A vector of Cox-Snell residuals created by the user based on the model being evaluated,
# 2. A status vector
# 3. Optional x- and y- limits for the resulting plot

CoxSnell = function(cs,status,xlim=NULL,ylim=NULL)
{
kmcs = survfit(Surv(jitter(cs,amount=(max(cs)-min(cs))/1000),status) ~ 1)$surv

plot(log(-log(kmcs)) ~ sort(log(cs)) ,
      xlab="log(Cox-Snell)", ylab="log(-log(S(Cox-Snell)))", xlim=xlim, ylim=ylim )

abline(0,1,col='red')
}
```

Log-normal model:

```
Dlnorm <- survreg(Surv(dtime, death) ~ Treatment + size + nodes + age , dist='lognormal', data=rc
Dlnorm
```

```
## Call:
## survreg(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##     age, data = rotterdam, dist = "lognormal")
##
## Coefficients:
##                   (Intercept)                TreatmentChemo
##                  9.709442268                  -0.431016329
##               TreatmentHormon TreatmentNaN/Other Treatment
##                  -0.346351742                  -0.423626557
##                     size20-50                       size>50
##                  -0.372703559                  -0.654189313
##                         nodes                           age
##                  -0.079103425                  -0.009903862
##
## Scale= 1.077329
##
## Loglik(model)= -12034.1   Loglik(intercept only)= -12286.5
##  Chisq= 504.67 on 7 degrees of freedom, p= <2e-16
## n= 2982
```

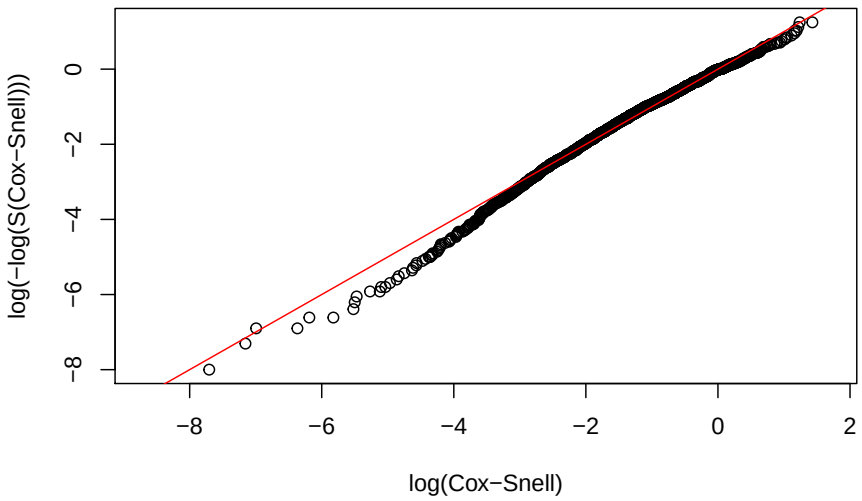
```
CS_Death <- -log(1 - plnorm(rotterdam$dtime, 9.709442268-0.431016329*(rotterdam$Treatme
                                                -0.346351742*(rotterdam$Treatment=="H
                                                -0.423626557*(rotterdam$Treatment=="N
                                                -0.372703559*(rotterdam$size=="20-50"
                                                -0.654189313*(rotterdam$size==">50")
                                                -0.079103425*rotterdam$nodes
                                                -0.009903862*rotterdam$age,
                                           1.077329))
# Make appropriate graph using CoxSnell function
CoxSnell(CS_Death, rotterdam$death)
```




```
Rlnorm <- survreg(Surv(rtime, recur) ~ Treatment + size + nodes + age, dist='lognormal', data=rot
Rlnorm
```

```
## Call:
## survreg(formula = Surv(rtime, recur) ~ Treatment + size + nodes +
##     age, data = rotterdam, dist = "lognormal")
##
## Coefficients:
##                   (Intercept)                TreatmentChemo
##                   8.514204484                  -0.382172447
##               TreatmentHormon TreatmentNaN/Other Treatment
##                  -0.479063703                  -0.605193863
##                     size20-50                       size>50
##                  -0.458345796                  -0.738657689
##                         nodes                           age
##                  -0.107708963                   0.009059467
##
## Scale= 1.340545
##
## Loglik(model)= -13803.8   Loglik(intercept only)= -14045.8
##  Chisq= 483.94 on 7 degrees of freedom, p= <2e-16
## n= 2982
```

```
CS_Recur <- -log(1 - plnorm(rotterdam$rtime, 8.514204484-0.382172447*(rotterdam$Treatme
                                                -0.479063703*(rotterdam$Treatment=="
                                                -0.605193863*(rotterdam$Treatment=="
                                                -0.458345796*(rotterdam$size=="20-50"
                                                -0.738657689*(rotterdam$size==">50")
                                                -0.107708963*rotterdam$nodes
                                                +0.009059467*rotterdam$age,
                                            1.340545))

# Make appropriate graph using CoxSnell function
CoxSnell(CS_Recur, rotterdam$recur)
```



3.4 Cox-PH model:

```
m_death = coxph(Surv(dtime, death) ~ Treatment + size + nodes + strata(age), data=rotte
m_death
```

```
## Call:
## coxph(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##     strata(age), data = rotterdam)
##
```

```
##                                  coef exp(coef) se(coef)      z        p
## TreatmentChemo               0.566734  1.762501 0.368984  1.536    0.125
## TreatmentHormon              0.571493  1.770909 0.375700  1.521    0.128
## TreatmentNaN/Other Treatment 0.497674  1.644890 0.365574  1.361    0.173
## size20-50                    0.464177  1.590704 0.066890  6.939 3.94e-12
## size>50                      0.772204  2.164532 0.097352  7.932 2.16e-15
## nodes                        0.078158  1.081294 0.005365 14.567  < 2e-16
##
## Likelihood ratio test=391.7  on 6 df, p=< 2.2e-16
## n= 2982, number of events= 1272
```

```
cox.zph(m_death)
```

```
##           chisq df    p
## Treatment  3.60  3 0.31
## size       4.32  2 0.12
## nodes      1.37  1 0.24
## GLOBAL     9.64  6 0.14
```

```
m_recur = coxph(Surv(rtime, recur) ~ size, data=rotterdam)
m_recur
```

```
## Call:
## coxph(formula = Surv(rtime, recur) ~ size, data = rotterdam)
##
##              coef exp(coef) se(coef)      z      p
## size20-50 0.55148   1.73583  0.05586  9.872 <2e-16
## size>50   1.04332   2.83861  0.08148 12.804 <2e-16
##
## Likelihood ratio test=185.3  on 2 df, p=< 2.2e-16
## n= 2982, number of events= 1518
```

```
cox.zph(m_recur)
```

```
##        chisq df       p
## size    35.5  2 1.9e-08
## GLOBAL  35.5  2 1.9e-08
```

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.