

Rotterdam

JACK TAN, YIMING MIAO

2021-03-09

Contents

1	Motivation	3
1.1	Some Background Information	3
2	Data Exploration	5
2.1	Loading Data	5
2.2	Data Wrangling	6
2.3	Data visualizations and exploration	7
3	Survival	20
3.1	Loading Data	20
3.2	Kaplan-Miere estimator of the entire dataset	20
3.3	Kaplan-Miere estimator on different variables in <code>rotterdam</code> . . .	23
3.4	Parametric Models	32
3.5	Cox-PH model:	40
4	Reference	43

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
```

The number of lymph nodes tested positive is a important measure in the TNM system. However, right now the `nodes` variable in our dataset is numeric and makes it hard for us to make visualization. For visualization purpose, we will make a new categorical variable called `Nodes_level`. For lymph nodes tested positive, the usual medical 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 method.

```
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))
```

Note that in this manner as we try to ‘merge’ two binary variables into one variable with four levels, we are assuming interaction between `chemo` and `hormon`.

```
rotterdam_recur <- rotterdam %>%
  filter(recur == 1) %>%
  mutate(drecurtime = dtime-rtime)
nrow(rotterdam_recur)
```

```
## [1] 1518
```

We also thought it would be of interest to investigate how recurrence of tumor might affect the survival of the patient. We made a new dataset called `rotterdam_recur`, which only include patients with `recur = 1`. Now the dataset contains 1518 data points, a little over the original `rotterdam` dataset. We will label the time from recurrence to death as `drecurtime` in the new data frame.

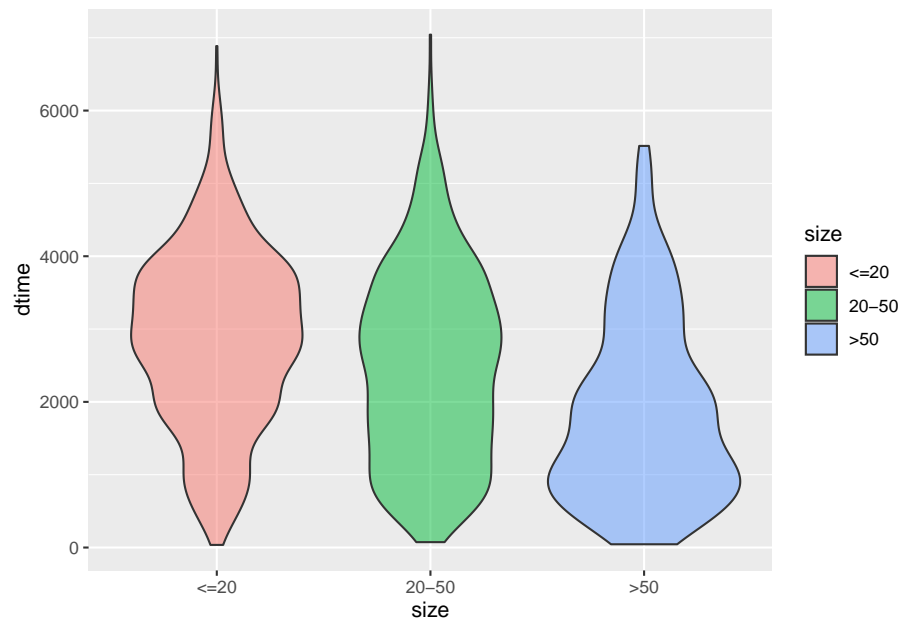
2.3 Data visualizations and exploration

2.3.1 Diagnostics vs. Survival Times

It is commonly considered that the earlier the breast cancer is detected and the earlier it is treated, the longer survival a patient might enjoy. Thus we think it is important to first look at the diagnostics before treatment and visualize their relationship with survival times.

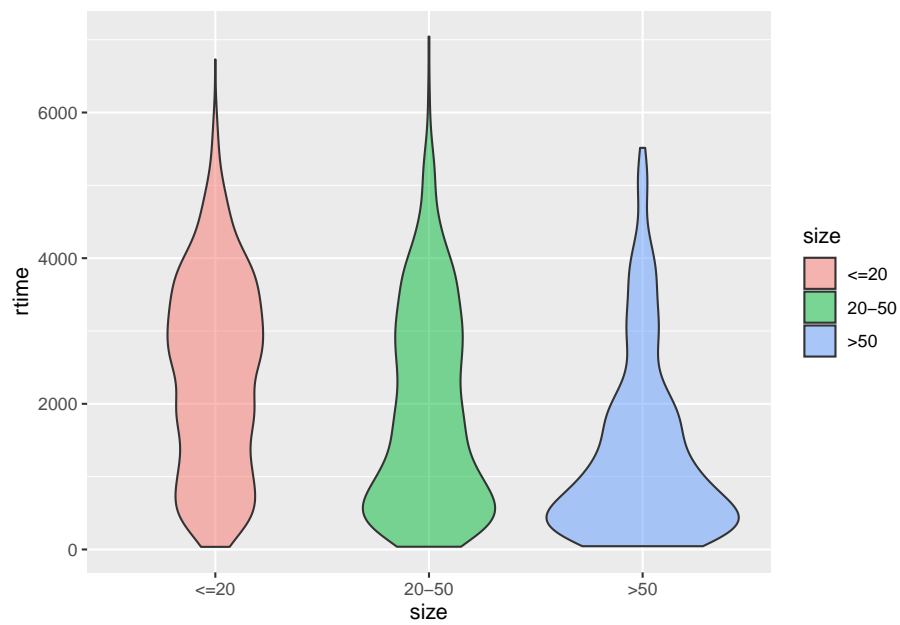
`size` vs. `dtime`

```
ggplot(data = rotterdam, aes(x = size, y = dtime, fill = size)) +  
  geom_violin(alpha = 0.5)
```



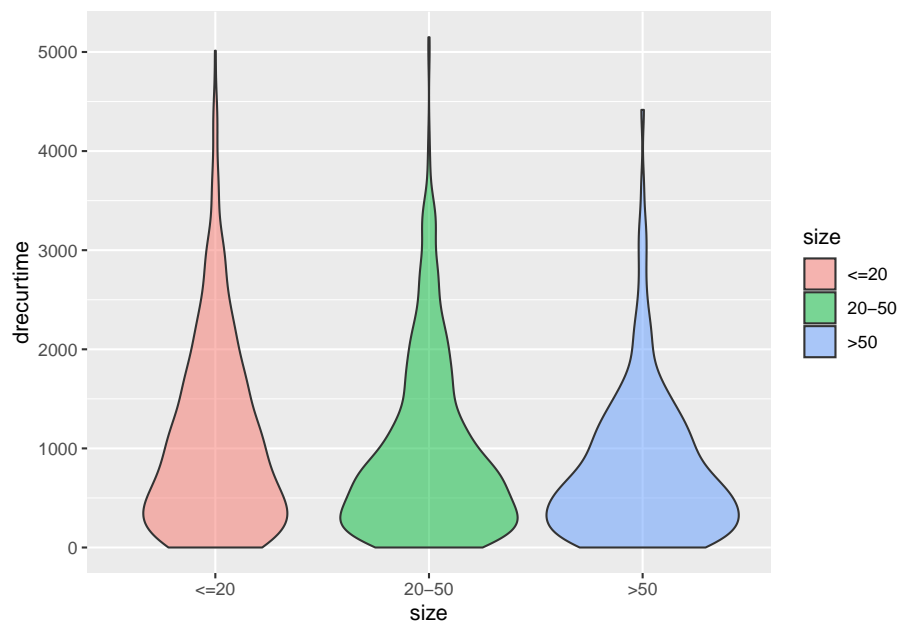
size vs. rtime

```
ggplot(data = rotterdam, aes(x = size, y = rtime, fill = size)) +  
  geom_violin(alpha = 0.5)
```

size vs. drecurtime

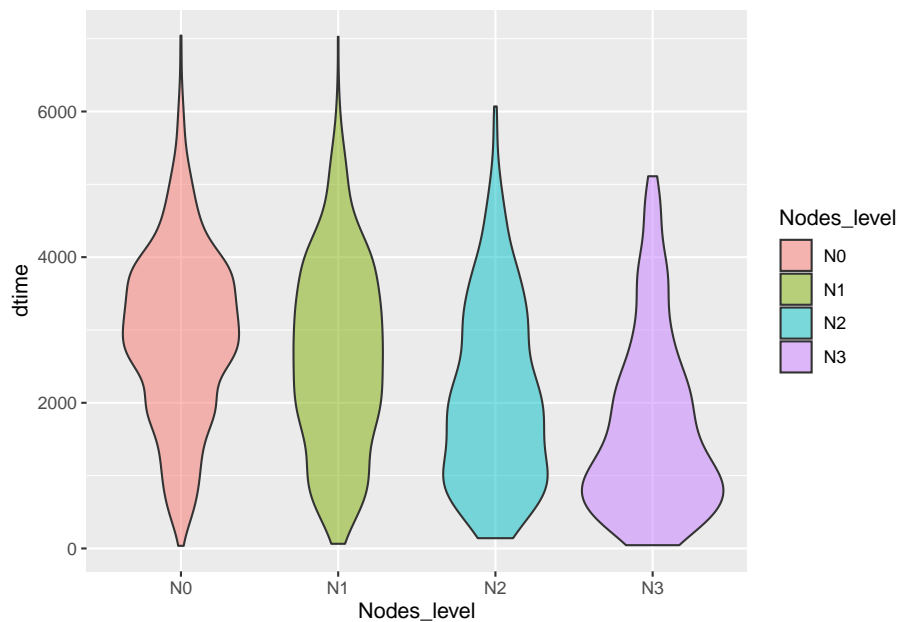
```
ggplot(data = rotterdam_recur, aes(x = size, y = drecurtime, fill = size)) +  
  geom_violin(alpha = 0.5)
```



As we can see from the three 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 cancer cells to recur in 500 days. However, after cancer cells have recurred, most patients could not survive over 2 years.

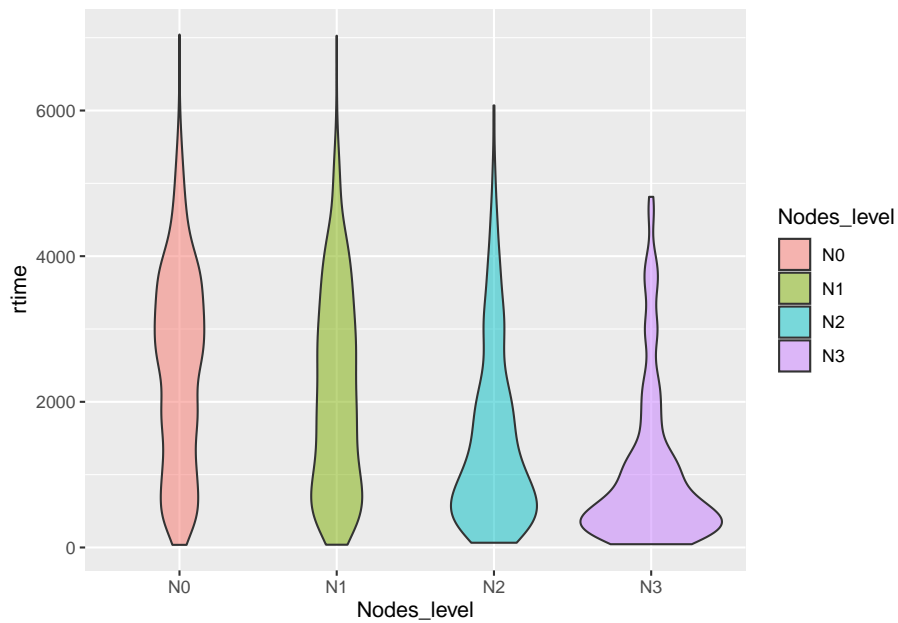
Nodes_level vs. dtime

```
ggplot(data = rotterdam, aes(x = Nodes_level, y = dtime, fill = Nodes_level)) +  
  geom_violin(alpha = 0.5)
```



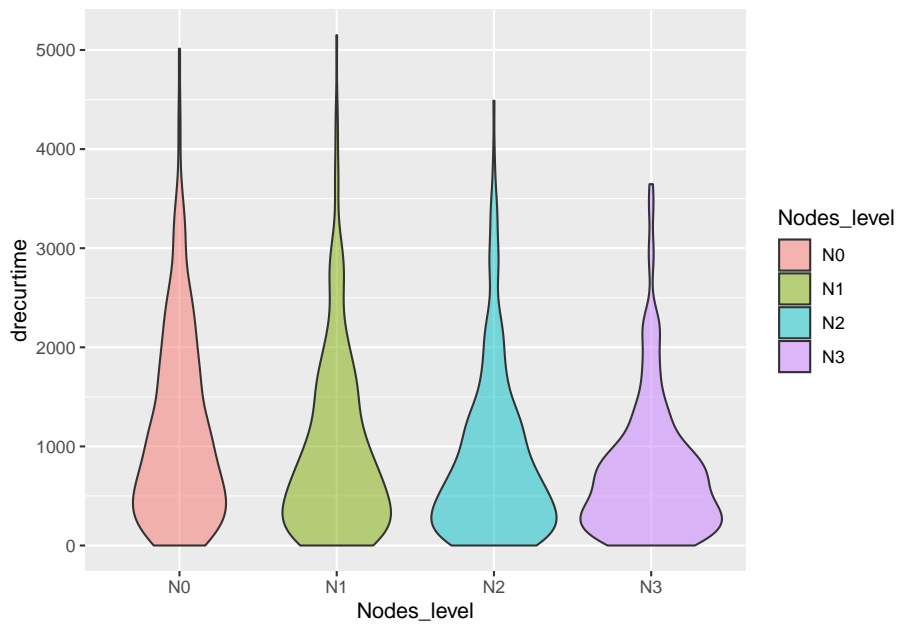
Nodes_level vs. rtime

```
ggplot(data = rotterdam, aes(x = Nodes_level, y = rtime, fill = Nodes_level)) +  
  geom_violin(alpha = 0.5)
```



Nodes_level vs. drecurtime

```
ggplot(data = rotterdam_recur, aes(x = Nodes_level, y = drecurtime, fill = Nodes_level)) +  
  geom_violin(alpha = 0.5)
```



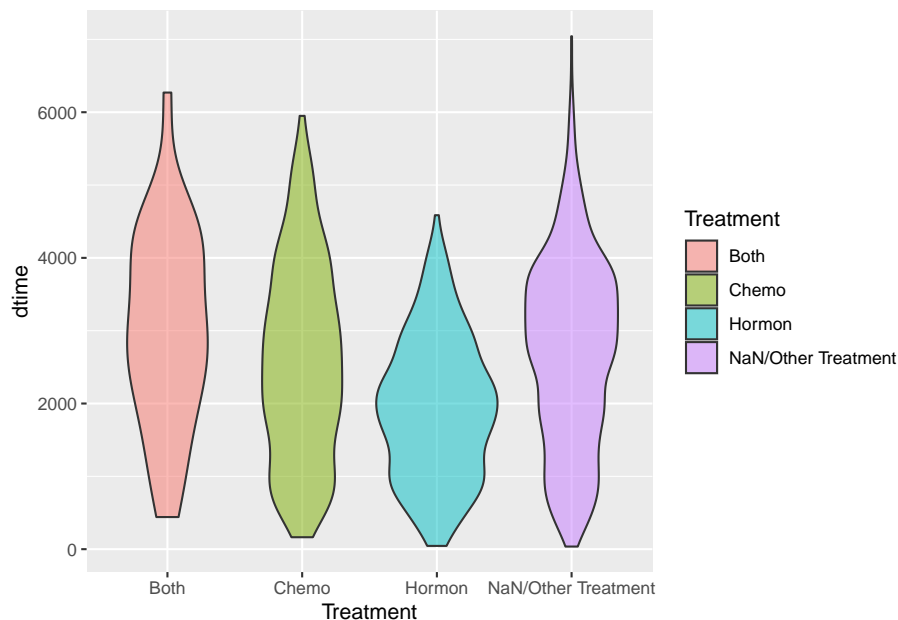
Similarly, `nodes` is also a factor impacting the life of breast cancer patients. In fact, for patients with high `Nodes_level`, it is typically considered they are either having metastasis of the cancer or already experiencing a regional recurrence of the cancer. Thus, we could see that most patients with N2 or N3 `Nodes_level` experience recurrence shortly after treatment. However, after tumor has recurred, most patients could not survive over 2 years.

2.3.2 Treatment vs. Survival Times

Next we are also going to look at the effect of different types of treatments on the survival times of breast cancer patients.

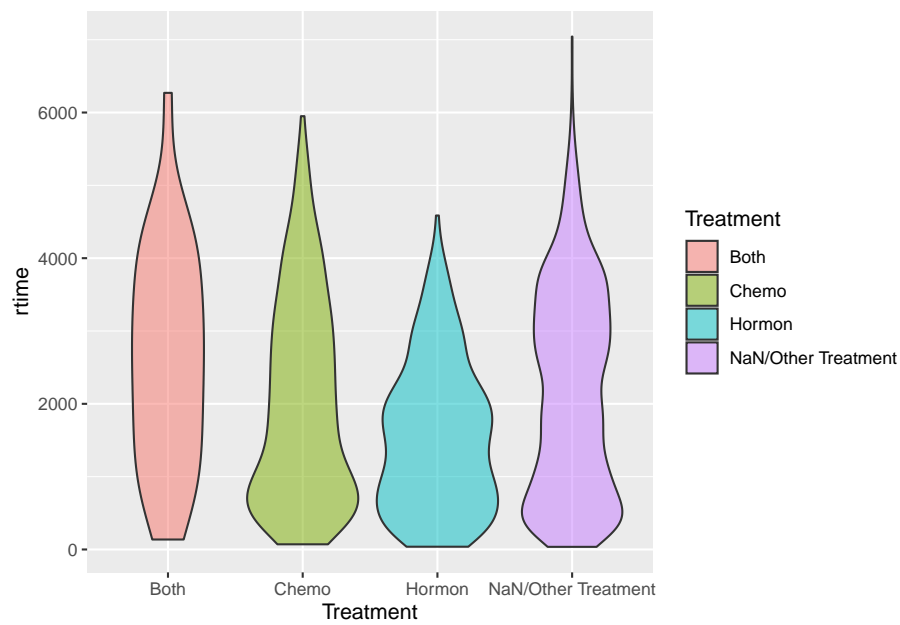
Treatment vs. `dttime`

```
ggplot(data = rotterdam, aes(x = Treatment, y = dttime, fill = Treatment)) +  
  geom_violin(alpha = 0.5)
```



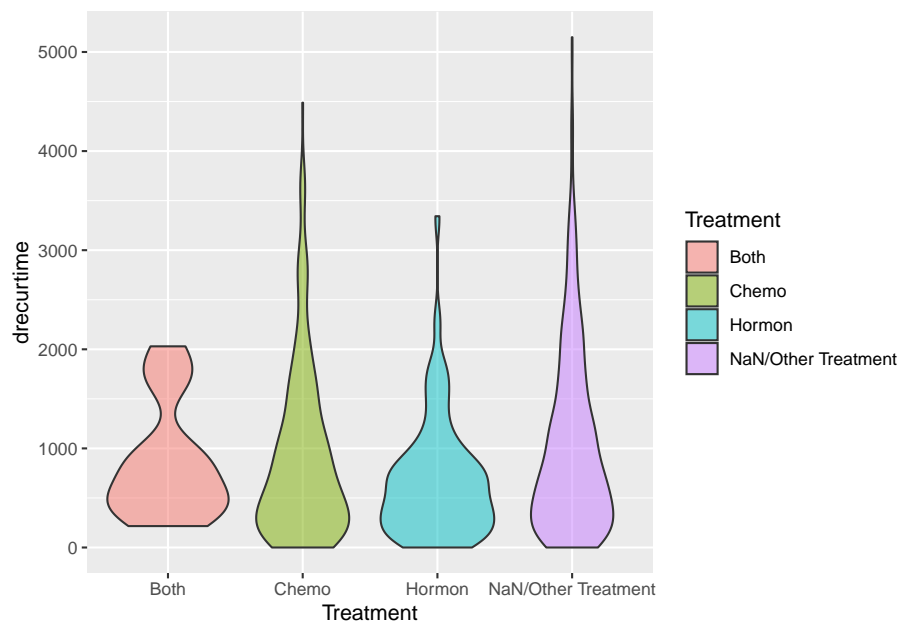
Treatment vs. `rtime`

```
ggplot(data = rotterdam, aes(x = Treatment, y = rtime, fill = Treatment)) +  
  geom_violin(alpha = 0.5)
```



Treatment vs. drecurtime

```
ggplot(data = rotterdam_recur, aes(x = Treatment, y = drecurtime, fill = Treatment)) +  
  geom_violin(alpha = 0.5)
```



By examining the three 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. However, after tumor have recurred, most patients do not live up to 2 years.

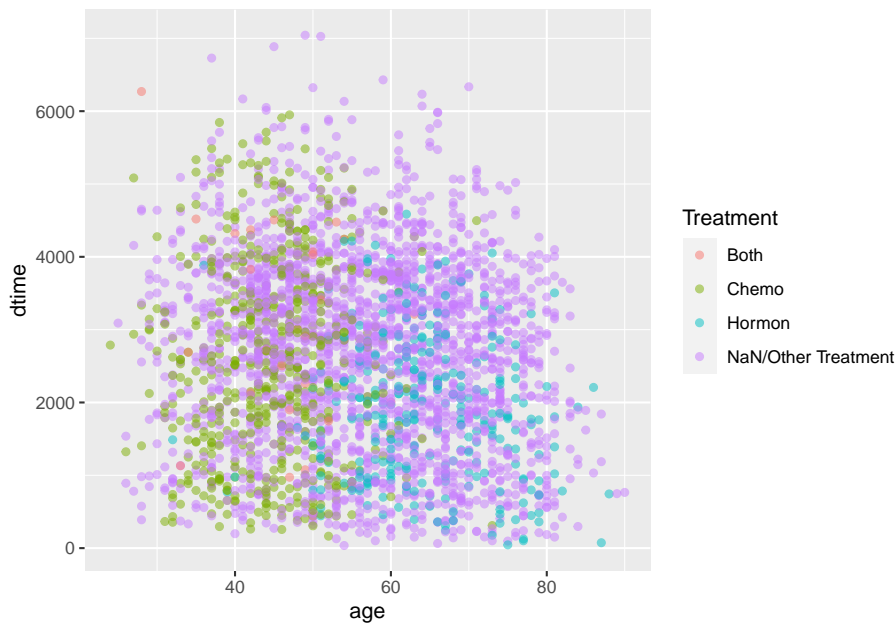
We also found a bi-modal shape in **NaN/Other Traetment** group in **rtime** vs. **Treatment**. This may because the two peaks correponds to no treatment and other treatment separately, but currently we don't have more information investigating the true reason.

2.3.3 age + Treatment vs. Survival Times

age + Treatment vs. dtime

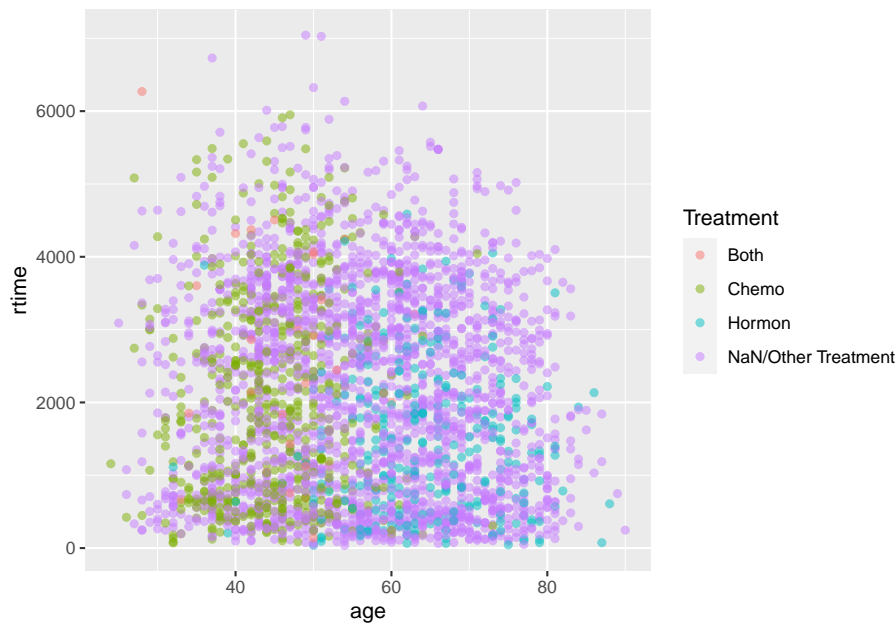
As we were visualizing for **Treatment** vs. **dtime**, we found that Hormontherapy generally has a weaker effect than Chemotherapy, but we think there might be some confounding variables that leads to such conclusion. One that we discovered is **age**:

```
ggplot(data = rotterdam, aes(x = age, y = dtime, color = Treatment)) +  
  geom_point(alpha = 0.5)
```



age + Treatment vs. rtime

```
ggplot(data = rotterdam, aes(x = age, y = rtime, color = Treatment)) +  
  geom_point(alpha = 0.5)
```

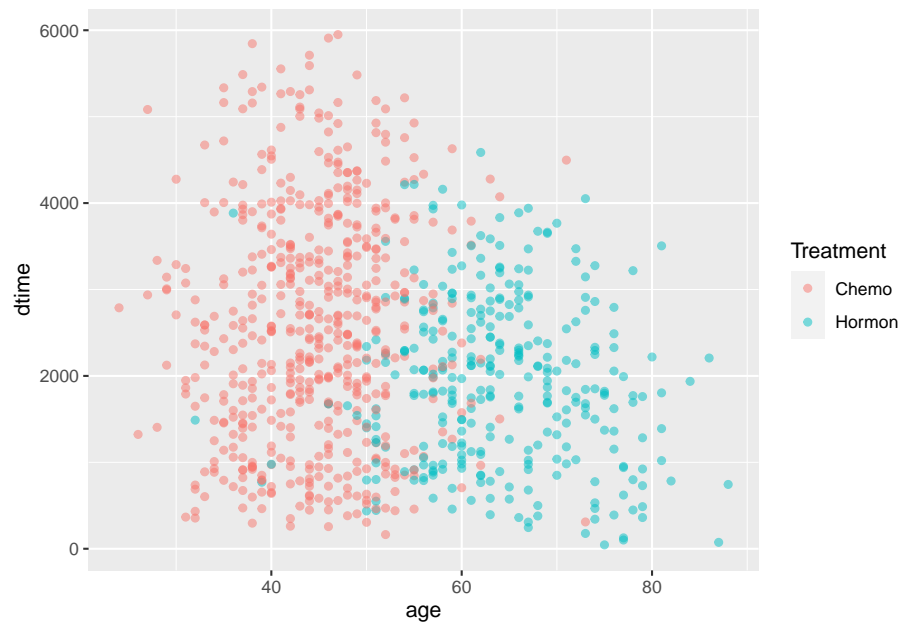


It might be different to see from the plots right now, so we decided to make a partial plot of the full plot by filtering the patients who did not take either treatment out.

```
rotterdam_new <- rotterdam %>%  
  filter(Treatment != "NaN/Other Treatment") %>%  
  filter(Treatment != "Both")
```

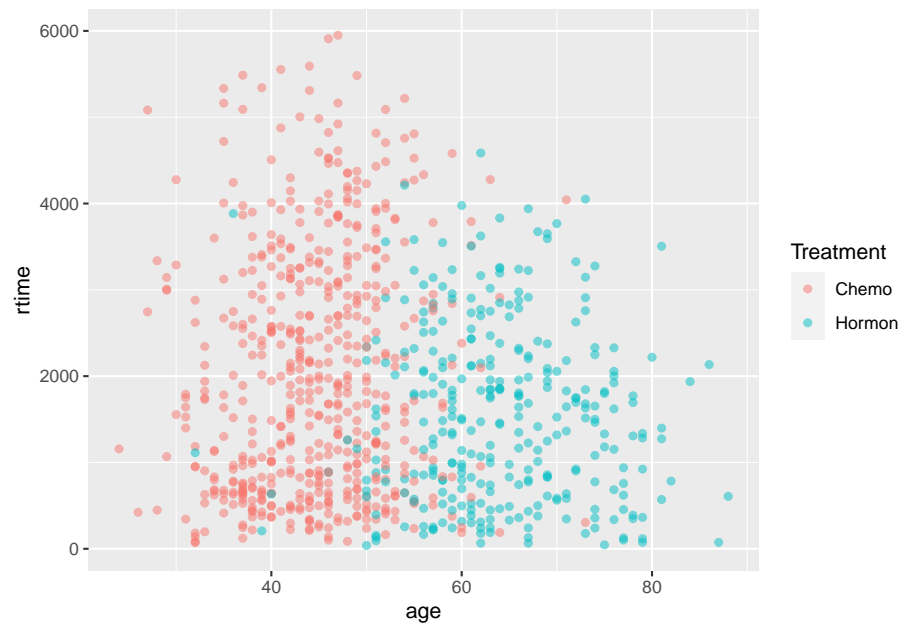
age + Treatment vs. dtime

```
ggplot(data = rotterdam_new, aes(x = age, y = dtime, color = Treatment)) +  
  geom_point(alpha = 0.5)
```



age + Treatment vs. rtime

```
ggplot(data = rotterdam_new, aes(x = age, y = rtime, color = Treatment)) +  
  geom_point(alpha = 0.5)
```



From the two plots above we could see that the patients who took either chemotherapy or hormone therapy are clearly clustered. For the group who only took chemotherapy, most patients' age are located below 50 years old. For the group who only took hormone therapy, most patients' age are located above 50 years old. This is because chemotherapy might have more negative effects for patients at larger age than hormone therapy and thus would effect survival if the wrong therapy is given. Generally, hormone therapy is more friendly to elder people but chemotherapy has better effect.

2.3.4 General X-year Survival Rate

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.

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

```
## # A tibble: 2 x 2
##   grade number
##   <fct>   <int>
## 1 2       794
## 2 3      2188
```

However, we have to notice that in our dataset, most patients are diagnosed with stage III breast cancer and a small portion are diagnosed with stage II breast cancer, which makes their cancer pretty severe already. This would give a pessimistic calculation of 5-year/ 10-year survival rates of breast cancer patients as a whole. In fact, according to webMD.com, the overall 5-year relative survival rate for breast cancer is 90% and the 10-year breast cancer relative survival rate is 84%.

Thus the important point is that female with high risk of breast cancer (family inheritance, bad life habits, etc.) should have regular physical examination, with proper screening for breast cancer. Even if diagnosed, do not panic and take treatment as soon as possible. In this way, a breast cancer patients might be able to enjoy longer survival.

Another important yet sad point is that after cancer has recurred, it does not matter what treatment a patient takes and most people do not live up to 2 years if recurred. Thus patients should be extremely careful not getting cancer recurred.

Chapter 3

Survival

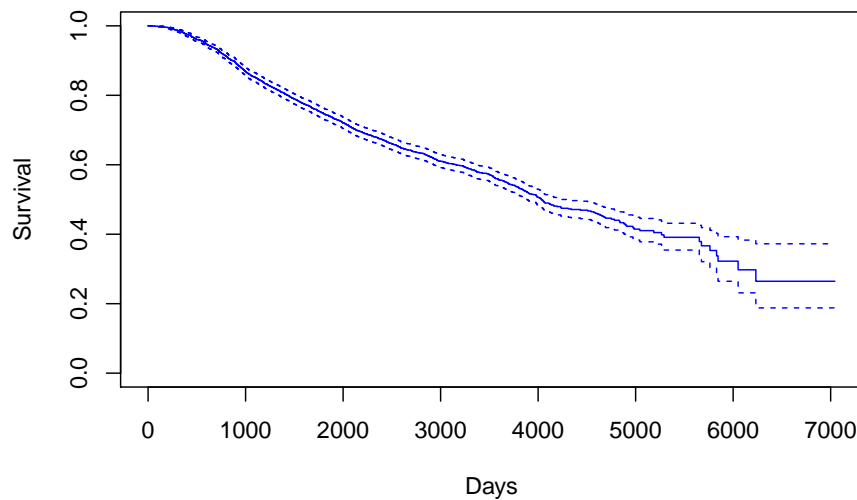
3.1 Loading Data

```
data(rotterdam)
```

3.2 Kaplan-Miere estimator of the entire dataset

Death Time

```
KM <- survfit(Surv(dtime, death) ~ 1, data = rotterdam)  
plot(KM, conf.int = TRUE, col = "blue", xlab="Days", ylab="Survival")
```



```
mean(rotterdam$dttime)
```

```
## [1] 2605.34
```

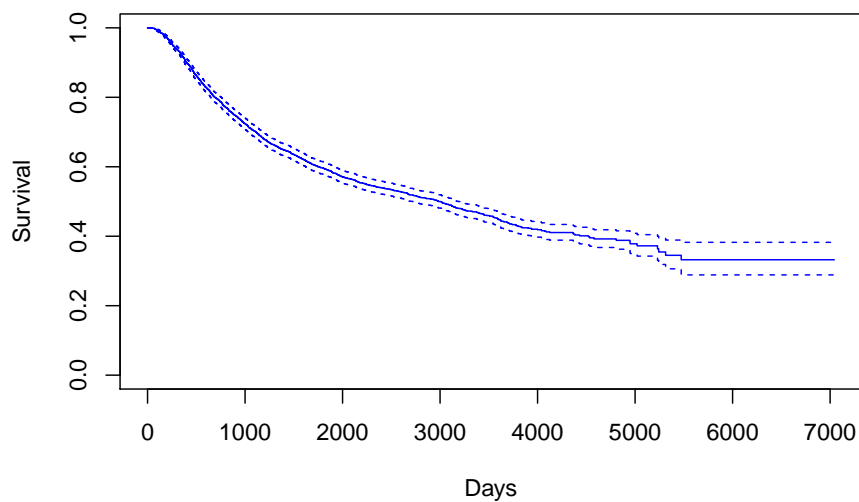
```
median(rotterdam$dttime)
```

```
## [1] 2638.5
```

The overall mean survival time till death for breast cancer is 2605 days, which is approximately 7 years. The overall median survival time till death for breast cancer is 2638 days, which is also approximately 7 years.

Recurrence Time

```
KM <- survfit(Surv(rtime, recur) ~ 1, data = rotterdam)
plot(KM, conf.int = TRUE, col = "blue", xlab="Days", ylab="Survival")
```



```
mean(rotterdam$rtime)
```

```
## [1] 2097.903
```

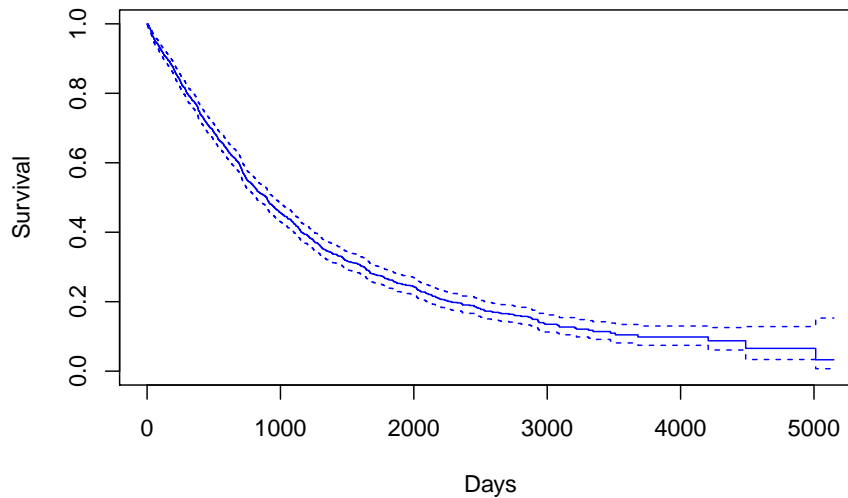
```
median(rotterdam$rtime)
```

```
## [1] 1940
```

The overall mean survival time till recurrence for breast cancer is 2097 days, which is approximately 6 years. The overall median survival time till recurrence for breast cancer is 1940 days, which is also approximately 5 years.

```
KM <- survfit(Surv(drecurtime, death) ~ 1, data = rotterdam_recur)
plot(KM, conf.int = TRUE, col = "blue", xlab="Days", ylab="Survival")
```

3.3. KAPLAN-MIERE ESTIMATOR ON DIFFERENT VARIABLES IN ROTTERDAM23



```
mean(rotterdam_recur$drecurtime)
```

```
## [1] 978.5481
```

```
median(rotterdam_recur$drecurtime)
```

```
## [1] 719.5
```

The overall mean survival time after recurrence till death for breast cancer is 834 days, which is approximately a little more than 2 years. The overall median survival time after recurrence till death for breast cancer is 625 days, which is approximately less than 2 years.

Since a Kaplan-Miere estimator is unbiased, we could view the median as being very close to the true value of survival time.

3.3 Kaplan-Miere estimator on different variables in rotterdam

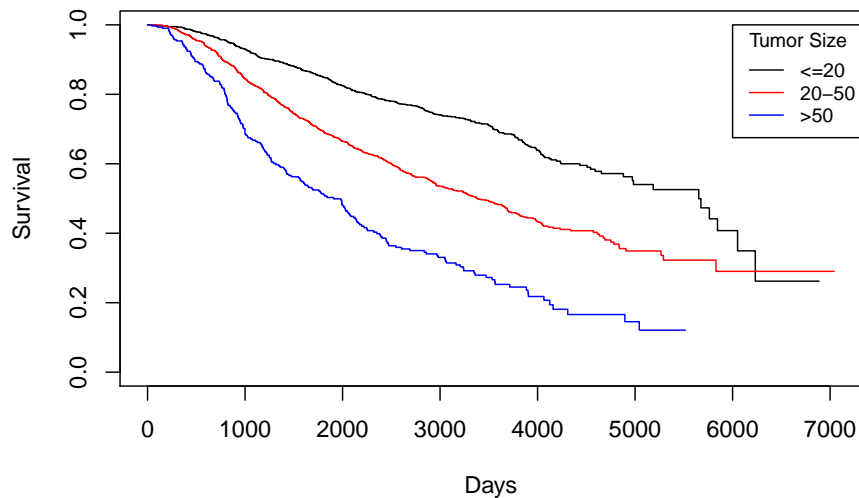
There are 16 variables in the `rotterdam` dataset. Of course we could have fit each variable with a KM estimator, but it would be meaningless to do them

all. We will stick to the Diagnostics and Treatment we mentioned in Chapter 2 and fit `size`, `Nodes_level` (we are not using nodes because a Kaplan-Meier estimator does not work well with quantitative variables), and `Treatment` each with KM estimators with respect to `dtime`, `rtime`, and `drecurtime` to grasp the survival time within each category of the variables.

3.3.1 size vs. Survival Times

size vs. dtime

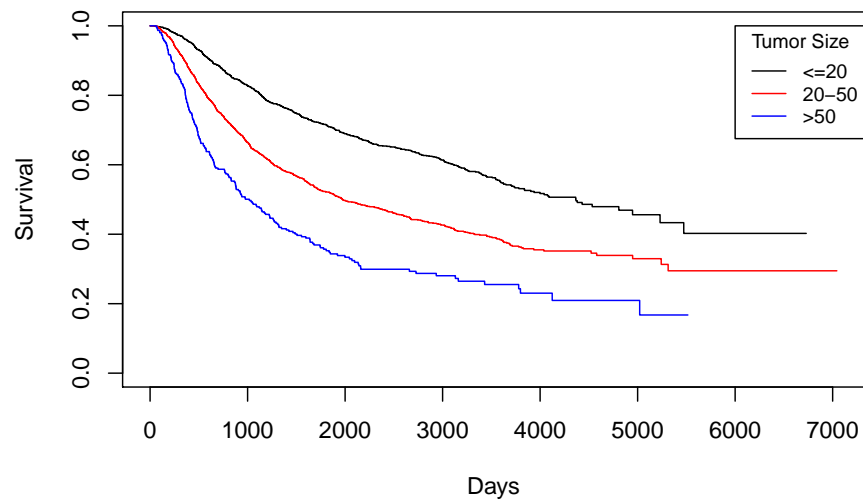
```
KM_None_Death <- survfit(Surv(dtime, death) ~ size, data = rotterdam)
plot(KM_None_Death, conf.type = "plain", col = c("black", "red", "blue"), xlab="Days", ylab="Survival",
     legend(6000, 1, legend=c("<=20", "20-50", ">50"),
           col=c("black", "red", "blue"), lty=1, cex=0.8,
         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



size vs. rtime

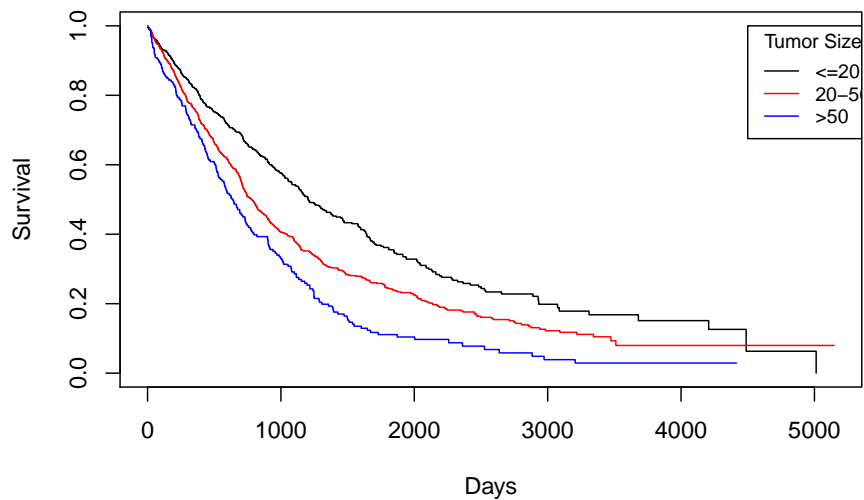
```
KM_None_Recur <- survfit(Surv(rtime, recur) ~ size, data = rotterdam)
plot(KM_None_Recur, conf.type = "plain", col = c("black", "red", "blue"), xlab="Days", ylab="Survival",
     legend(6000, 1, legend=c("<=20", "20-50", ">50"),
           col=c("black", "red", "blue"), lty=1, cex=0.8,
         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


3.3. KAPLAN-MIERE ESTIMATOR ON DIFFERENT VARIABLES IN ROTTERDAM25



In general, patients with smaller tumor at diagnosis enjoys longer survival for both death and recurrence.

```
KM_None_drecur <- survfit(Surv(drecurtime, death) ~ size, data = rotterdam_recur)
plot(KM_None_drecur, conf.type = "plain", col = c("black", "red", "blue"), xlab="Days", ylab="Survival",
     legend(4500, 1, legend=c("<=20", "20-50", ">50"),
          col=c("black", "red", "blue"), lty=1, cex=0.8,
          title="Tumor Size", text.font=6))
```



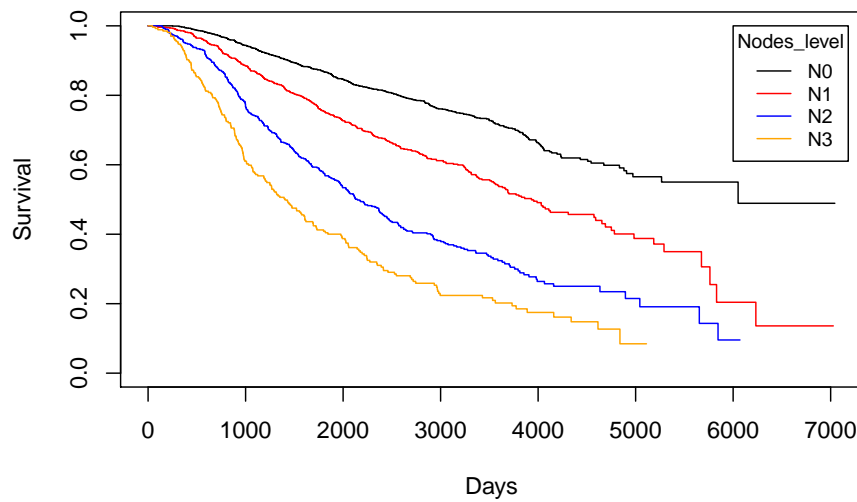
The trend is still the same as patients with smaller tumor size enjoy longer survival of death after recurrence, but the survival time now decreases much faster for all groups.

3.3.2 Nodes_level vs. Survival Times

Nodes_level vs. dttime

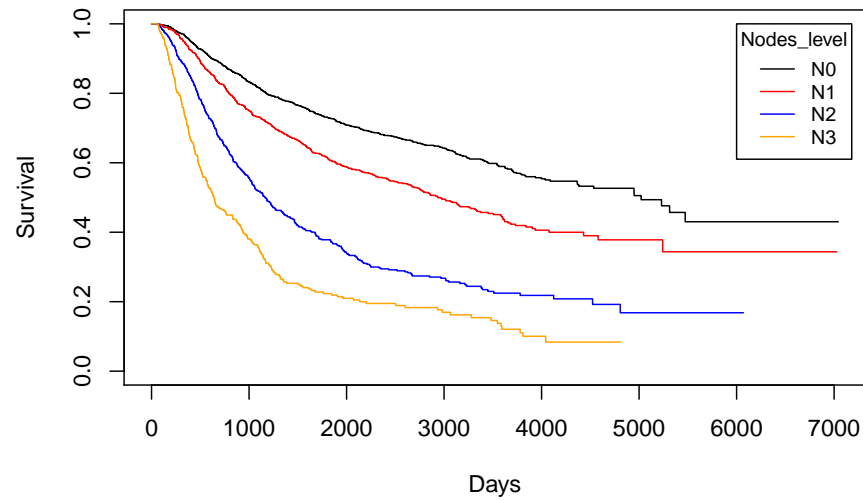
```
KM_None_Death <- survfit(Surv(dttime, death) ~ Nodes_level, data = rotterdam)
plot(KM_None_Death, conf.type = "plain", col = c("black", "red", "blue", "orange"), xlab=
legend(6000, 1, legend=c("N0", "N1", "N2", "N3"),
      col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,
      title="Nodes_level", text.font=6)
```

3.3. KAPLAN-MIERE ESTIMATOR ON DIFFERENT VARIABLES IN ROTTERDAM27



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="Days", ylab="Survival",
     legend=6000, 1, legend=c("N0", "N1", "N2", "N3"),
     col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,
     title="Nodes_level", text.font=6)
```

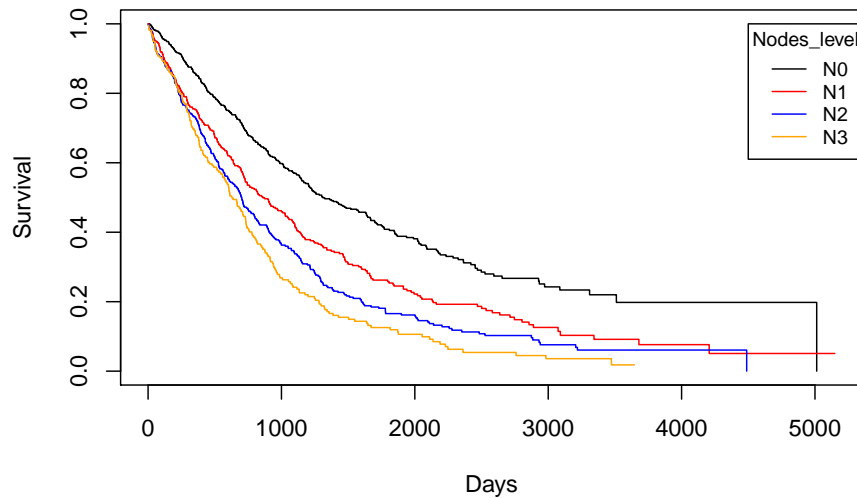


In general, patients with less nodes tested positive will enjoy longer survival for both death and recurrence.

Nodes_level vs. drecurtime

```
KM_None_drecur <- survfit(Surv(drecurtime, death) ~ Nodes_level, data = rotterdam_recur)
plot(KM_None_drecur, conf.type = "plain", col = c("black", "red", "blue", "orange"), xlab = "Days",
     legend(4500, 1, legend=c("N0", "N1", "N2", "N3"),
     col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,
     title="Nodes_level", text.font=6)
```

3.3. KAPLAN-MIERE ESTIMATOR ON DIFFERENT VARIABLES IN ROTTERDAM29

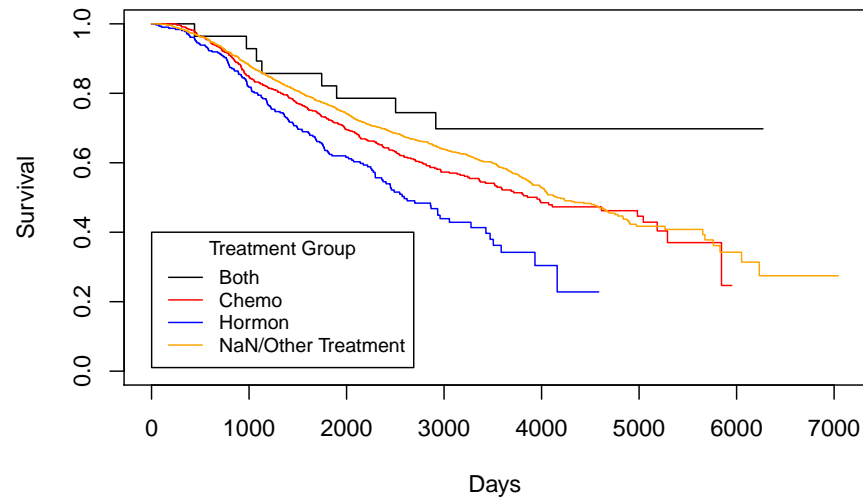


The trend is still the same as patients with fewer nodes tested positive enjoy longer survival of death after recurrence, but the survival time now decreases much faster for all groups, and the difference is small in groups N1, N2 and N3.

3.3.3 Treatment vs. Survival Times

Treatment vs. dttime

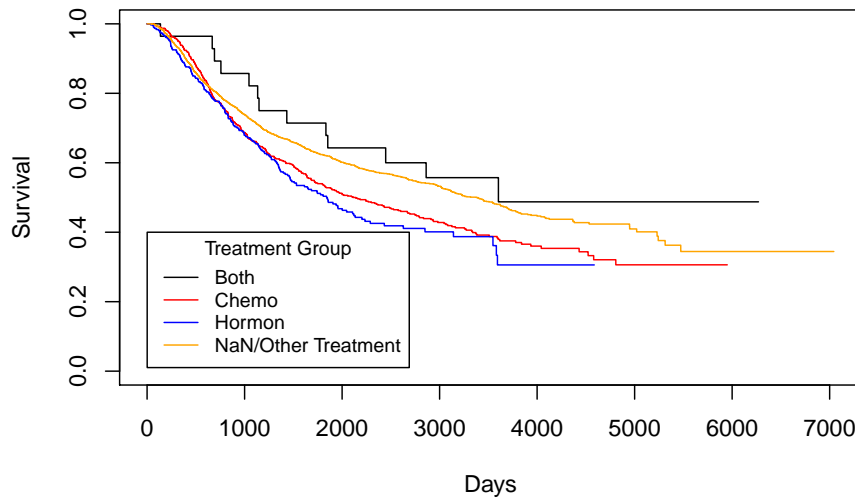
```
KM_Treatment_Death <- survfit(Surv(dttime, death) ~ Treatment, data = rotterdam)
plot(KM_Treatment_Death, conf.int = FALSE, col = c("black", "red", "blue", "orange"), xlab="Days",
legend(1, 0.4, legend=c("Both", "Chemo", "Hormon", "NaN/Other Treatment"),
col=c("black", "red", "blue", "orange"), lty=1, 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"),
     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))
```

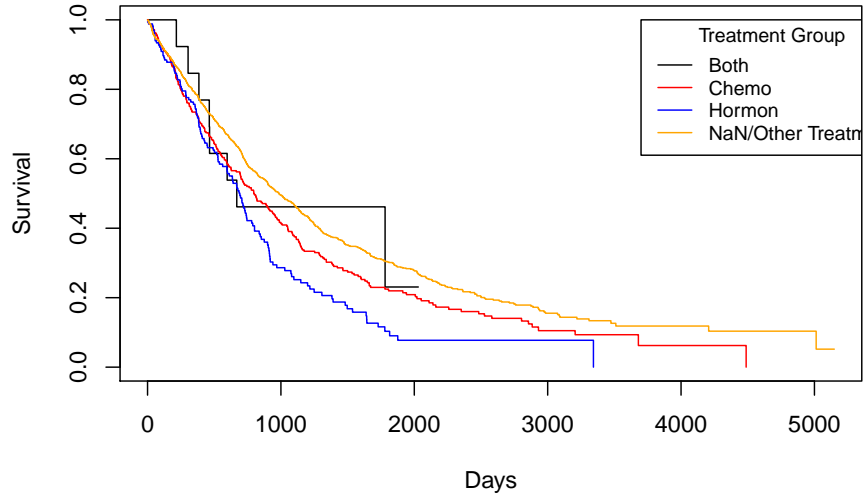
3.3. KAPLAN-MIERE ESTIMATOR ON DIFFERENT VARIABLES IN ROTTERDAM31



As we have discussed in Chapter 2, we know that generally chemotherapy is used on patients with age lower than 50 years old and hormonotherapy is used on patients with age higher than 50 years old. Based on the difference of treatment, we could see that chemotherapy has a better effect than hormonotherapy with respect to death time and a smaller yet still better effect regarding the recurrence time.

Treatment vs. drecurtime

```
KM_Treatment_drecur <- survfit(Surv(drecurtime, death) ~ Treatment, data = rotterdam_recur)
plot(KM_Treatment_drecur, conf.int = FALSE, col = c("black", "red", "blue", "orange"), xlab="Days",
     legend(3700, 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))
```



However, sadly enough, from the above plot we could see that no matter what treatment a patient use, it does not make a difference for the death survival time after cancer cells have recurred. This matches our conclusion from the visualization in Chapter 2.

3.4 Parametric Models

Another point that we are going to explore is if we would be able to fit our data to a parametric model. This matters since if we could fit any parametric model, then we should have a model good enough to generate predictions of breast cancer patients' survival and would have nice and interpretable coefficients to work with.

To do so, we will begin by checking if any of Exponential, Weibull, or Log-normal distribution would be adequate parametric assumption to cast on our data. We will verify the adequacy by checking the Cox-Snell residual plot. We will be fitting models using variables: **Treatment**, **size**, **nodes**, **age** (We have shown in Chapter 2 that age is a confounder for categories in Treatment).

3.4.1 User-defined Cox-Snell function


```

# 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')
}

```

3.4.2 Exponential models

We will begin by verifying the adequacy of Exponential model.

```

Dexp <- survreg(Surv(dtime, death) ~ Treatment + size + nodes + age, dist='exponential', data=rot)
Dexp

```

```

## Call:
## survreg(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##      age, data = rotterdam, dist = "exponential")
##
## Coefficients:
##              (Intercept)              TreatmentChemo
##              10.49678202              -0.51997169
##      TreatmentHormon TreatmentNaN/Other Treatment
##              -0.44961578              -0.46574886
##              size20-50              size>50
##              -0.46803053              -0.81548193
##              nodes              age
##              -0.06997150              -0.01412165
##
## Scale fixed at 1
##
## Loglik(model)= -12137   Loglik(intercept only)= -12360.4
##  Chisq= 446.94 on 7 degrees of freedom, p= <2e-16
## n= 2982

```

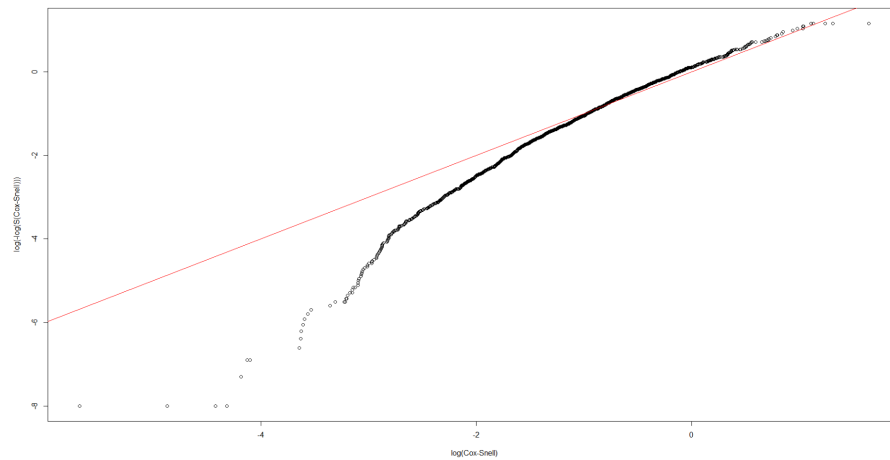


Figure 3.1:

```
Dexp <- survreg(Surv(rtime, recur) ~ Treatment + size + nodes + age, dist='exponential')
Dexp
```

```
## Call:
## survreg(formula = Surv(rtime, recur) ~ Treatment + size + nodes +
##      age, data = rotterdam, dist = "exponential")
##
## Coefficients:
##              (Intercept)              TreatmentChemo
##              8.818707424              -0.372797596
##      TreatmentHormon TreatmentNaN/Other Treatment
##      -0.536512833              -0.490056002
##      size20-50              size>50
##      -0.404626936              -0.719554358
##      nodes              age
##      -0.082728526              0.008130282
##
## Scale fixed at 1
##
## Loglik(model)= -13917.5   Loglik(intercept only)= -14153.7
##  Chisq= 472.4 on 7 degrees of freedom, p= <2e-16
## n= 2982
```

The Exponential does not seem adequate in this case.

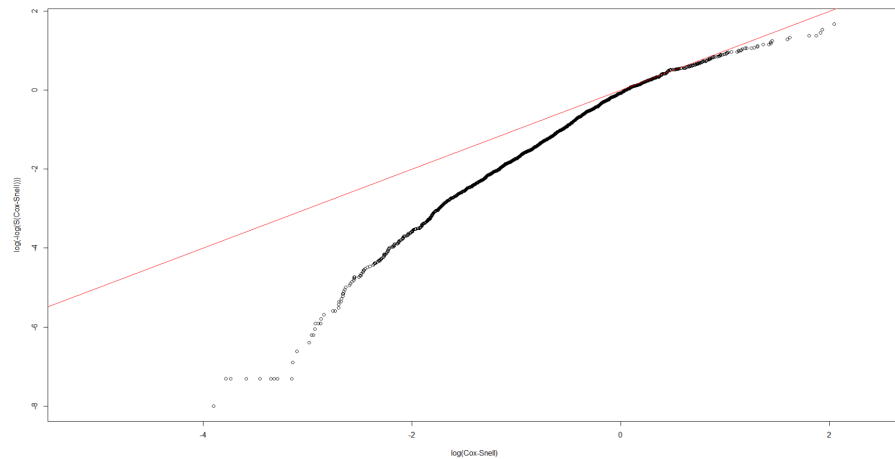


Figure 3.2:

3.4.3 Weibull models

Next we will examine the adequacy of Weibull model.

```
Dweibull <- survreg(Surv(dtime, death) ~ Treatment + size + nodes + age, dist='weibull', data=rotterdam)
Dweibull
```

```
## Call:
## survreg(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##   age, data = rotterdam, dist = "weibull")
##
## Coefficients:
##              (Intercept)              TreatmentChemo
##              9.93741760              -0.40097983
##      TreatmentHormon TreatmentNaN/Other Treatment
##              -0.40778590              -0.36210464
##              size20-50              size>50
##              -0.35546640              -0.65448317
##              nodes              age
##              -0.05589248              -0.01125732
##
## Scale= 0.739963
##
## Loglik(model)= -12068.9   Loglik(intercept only)= -12322.7
##   Chisq= 507.49 on 7 degrees of freedom, p= <2e-16
##   n= 2982
```

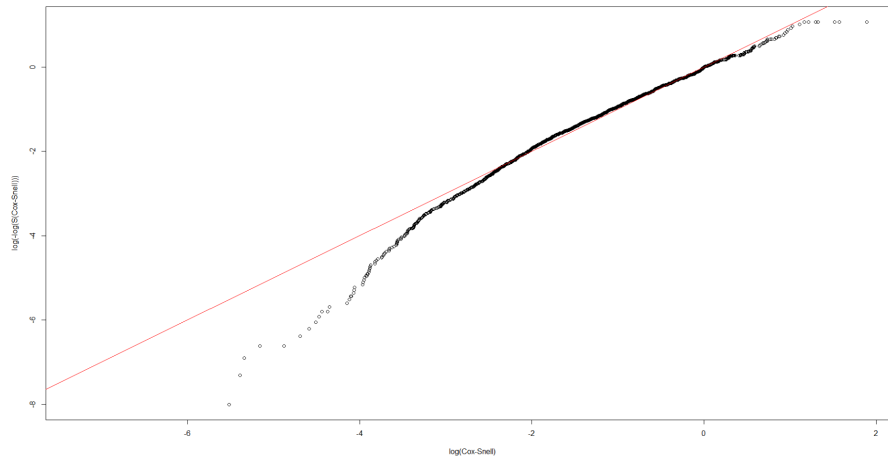


Figure 3.3:

```
Dweibull <- survreg(Surv(rtime, recur) ~ Treatment + size + nodes + age, dist='weibull')
Dweibull
```

```
## Call:
## survreg(formula = Surv(rtime, recur) ~ Treatment + size + nodes +
##   age, data = rotterdam, dist = "weibull")
##
## Coefficients:
##              (Intercept)              TreatmentChemo
##              8.830213089              -0.376480833
##      TreatmentHormon TreatmentNaN/Other Treatment
##      -0.539278367              -0.495766112
##           size20-50              size>50
##      -0.410910817              -0.727355688
##           nodes              age
##      -0.083786844              0.008292188
##
## Scale= 1.018498
##
## Loglik(model)= -13917.1   Loglik(intercept only)= -14145.7
##   Chisq= 457.19 on 7 degrees of freedom, p= <2e-16
## n= 2982
```

The Weibull model is still not adequate.

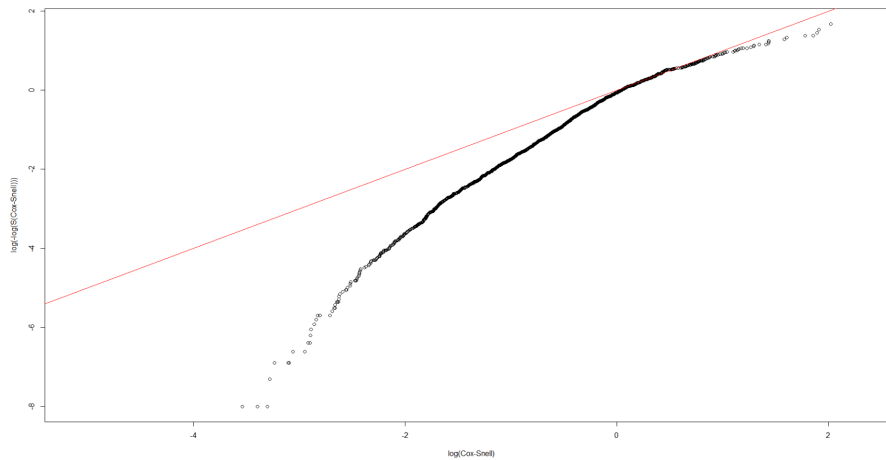


Figure 3.4:

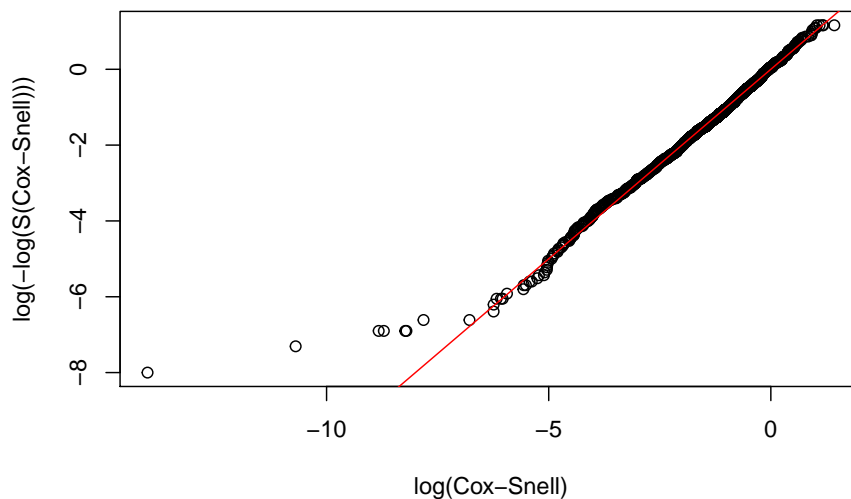
3.4.4 Log-normal models

Finally, we will examine the adequacy of Log-normal model.

```
Dlnorm <- survreg(Surv(dtime, death) ~ Treatment + size + nodes + age, dist='lognormal', data=rotterdam)
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_LnormD <- -log(1 - plnorm(rotterdam$ftime, 9.709442268-0.431016329*(rotterdam$Treatment=="1"
-0.346351742*(rotterdam$Treatment=="1"
-0.423626557*(rotterdam$Treatment=="1"
-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_LnormD, rotterdam$death)
```



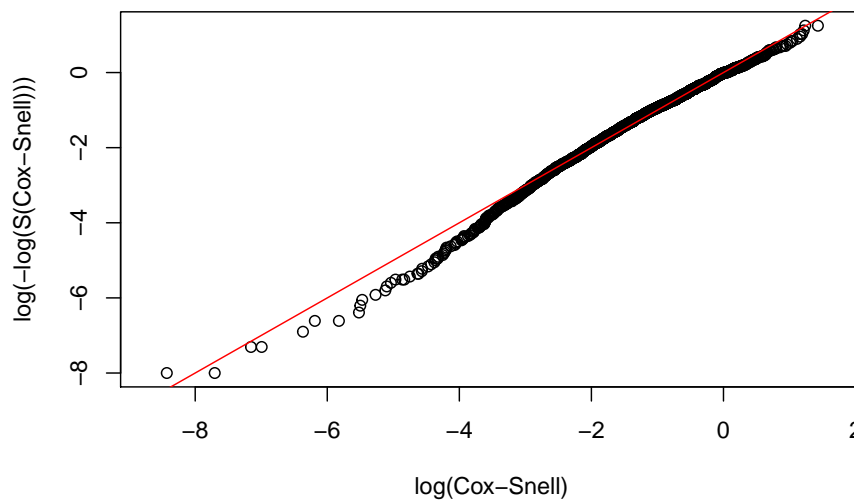
```
Rlnorm <- survreg(Surv(rtime, recur) ~ Treatment + size + nodes + age, dist='lognormal')
Rlnorm
```

```
## 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_LnormR <- -log(1 - plnorm(rotterdam$rtime, 8.514204484-0.382172447*(rotterdam$Treatment=="Chem
                                -0.479063703*(rotterdam$Treatment=="Hormon")
                                -0.605193863*(rotterdam$Treatment=="NaN/Other T
                                -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_LnormR, rotterdam$recur)
```



We could see that the Log-normal parametric model is an adequate model for both the `dttime` and `rtime` vs. `Treatment` + `size` + `nodes` + `age`.

TODO: why log-normal fit.

TODO: explain why age in `rtime` models is positive: competing risks.

3.5 Cox-PH model:

```
m_death_withage = coxph(Surv(dtime, death) ~ Treatment + size + nodes + age, data=rotterdam)
m_death_withage
```

```
## Call:
## coxph(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##       age, data = rotterdam)
##
##               coef exp(coef) se(coef)      z      p
## TreatmentChemo    0.546830  1.727768 0.360076  1.519  0.129
## TreatmentHormon    0.519838  1.681756 0.366212  1.420  0.156
## TreatmentNaN/Other Treatment 0.496324  1.642672 0.356927  1.391  0.164
## size20-50          0.477235  1.611612 0.065146  7.326 2.38e-13
## size>50            0.865167  2.375402 0.090893  9.519 < 2e-16
## nodes              0.074146  1.076964 0.004864 15.244 < 2e-16
## age                0.014904  1.015016 0.002562  5.818 5.94e-09
##
## Likelihood ratio test=487.1 on 7 df, p=< 2.2e-16
## n= 2982, number of events= 1272
```

```
cox.zph(m_death_withage)
```

```
##           chisq df      p
## Treatment   4.45  3 0.21657
## size        4.81  2 0.09023
## nodes       3.32  1 0.06850
## age        15.12  1 0.00010
## GLOBAL     25.38  7 0.00065
```

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

```
## 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_strataage)
```

```
##          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) ~ Treatment + size + nodes + age, data=rotterdam)
m_recur
```

```
## Call:
## coxph(formula = Surv(rtime, recur) ~ Treatment + size + nodes +
##       age, data = rotterdam)
##
##              coef exp(coef) se(coef)      z      p
## TreatmentChemo    0.324447  1.383266 0.283541  1.144 0.252512
## TreatmentHormon    0.436864  1.547846 0.290953  1.501 0.133227
## TreatmentNaN/Other Treatment 0.459422  1.583159 0.280366  1.639 0.101286
## size20-50          0.399867  1.491626 0.057857  6.911 4.80e-12
## size>50            0.684610  1.982999 0.087589  7.816 5.44e-15
## nodes              0.080239  1.083546 0.004575 17.539 < 2e-16
## age                -0.008653  0.991385 0.002314 -3.739 0.000184
##
## Likelihood ratio test=429.7 on 7 df, p=< 2.2e-16
## n= 2982, number of events= 1518
```

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

```
##          chisq df    p
## Treatment  7.943  3 0.047
## size      26.427  2 1.8e-06
## nodes      5.297  1 0.021
## age        0.104  1 0.747
## GLOBAL    42.374  7 4.4e-07
```

```
m_recur = coxph(Surv(rtime, recur) ~ Treatment + size + nodes + strata(age), data=rotterdam)
m_recur
```

```
## Call:
## coxph(formula = Surv(rtime, recur) ~ Treatment + size + nodes +
##       strata(age), data = rotterdam)
##
##               coef exp(coef) se(coef)      z      p
## TreatmentChemo    0.248511  1.282115 0.287150  0.865  0.387
## TreatmentHormon    0.294341  1.342241 0.295465  0.996  0.319
## TreatmentNaN/Other Treatment 0.352181  1.422166 0.284206  1.239  0.215
## size20-50          0.400691  1.492855 0.058893  6.804 1.02e-11
## size>50            0.684020  1.981830 0.091781  7.453 9.14e-14
## nodes              0.082436  1.085930 0.005002 16.482 < 2e-16
##
## Likelihood ratio test=411.3 on 6 df, p=< 2.2e-16
## n= 2982, number of events= 1518
```

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

```
##           chisq df      p
## Treatment   7.29  3  0.063
## size       21.99  2 1.7e-05
## nodes        6.32  1  0.012
## GLOBAL     37.04  6 1.7e-06
```

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.
- Martin, Laura J. “Breast Cancer: What Are the Survival Rates?” WebMD, WebMD, 13 May 2020, www.webmd.com/breast-cancer/guide/breast-cancer-survival-rates#:~:text=The 10-year breast cancer,are alive after 10 years)).