Rotterdam

JACK TAN, YIMING MIAO

2021-03-12

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 8

3 Survival **22**

3.1 Loading Data 22

3.2 Kaplan-Miere estimator of the entire dataset 22

3.3 Kaplan-Miere estimator on different variables in `rotterdam` . . . 25

3.4 Parametric Models 34

3.5 Cox-PH model: 43

4 Conclusion **47**

4.1 Limitation 47

4.2 Reference 48

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 of breast cancer. 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 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

2.2.1 T(Tumor)N(Node)M(Metastasis)

```
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))
```

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

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

In our dataset, we could see that all of the breast cancer patients are located at stage II or stage III breast cancer, which means that the cancer has not started shifting to other parts of the body, which makes the recordance of metastasis information to be impossible. As we will continue using the TNM system mentioned in the previous chapter to conduct our research, we won't be able to measure statistics related to metastasis.

2.2.2 Treatment

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 invesigate how recurrnce 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.

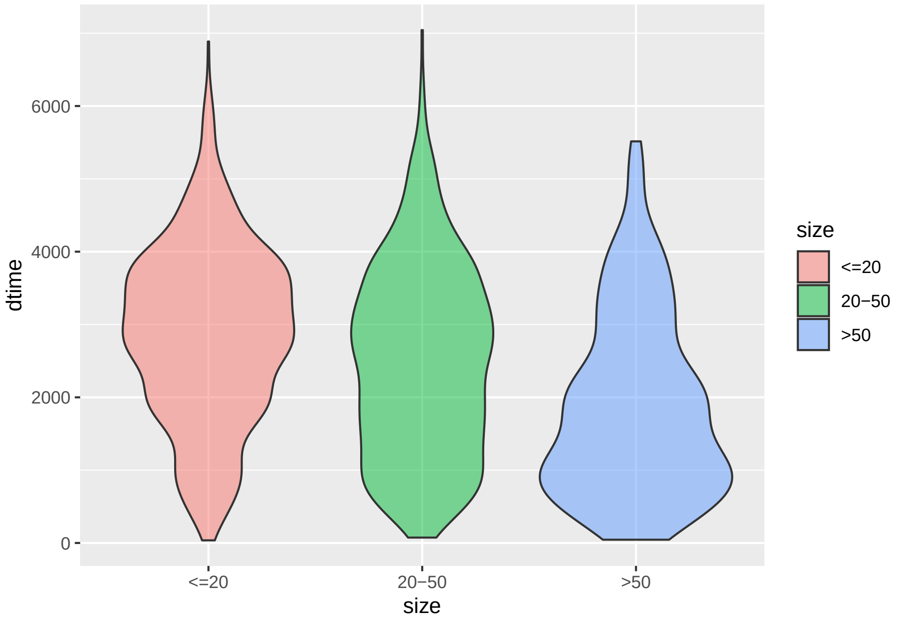
2.3 Data visualizations and exploration

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.

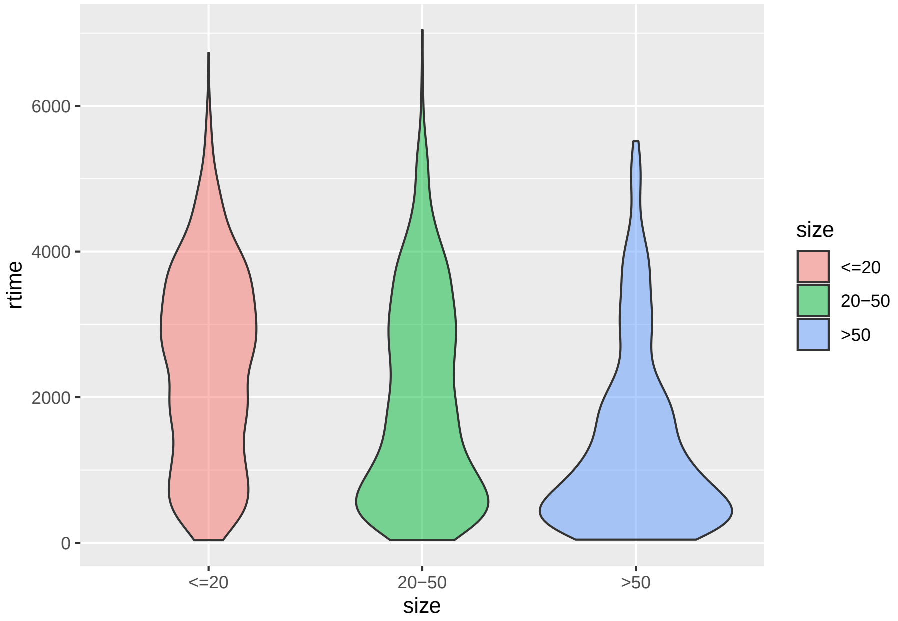
`size` vs. `dtime`

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



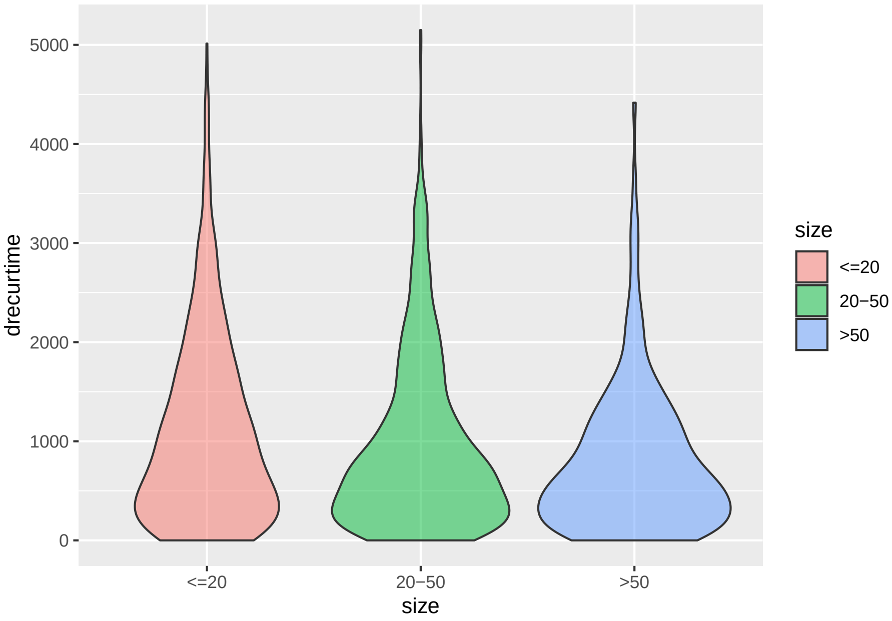
`size` vs. `rtime`

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



`size` vs. `drecurtime`

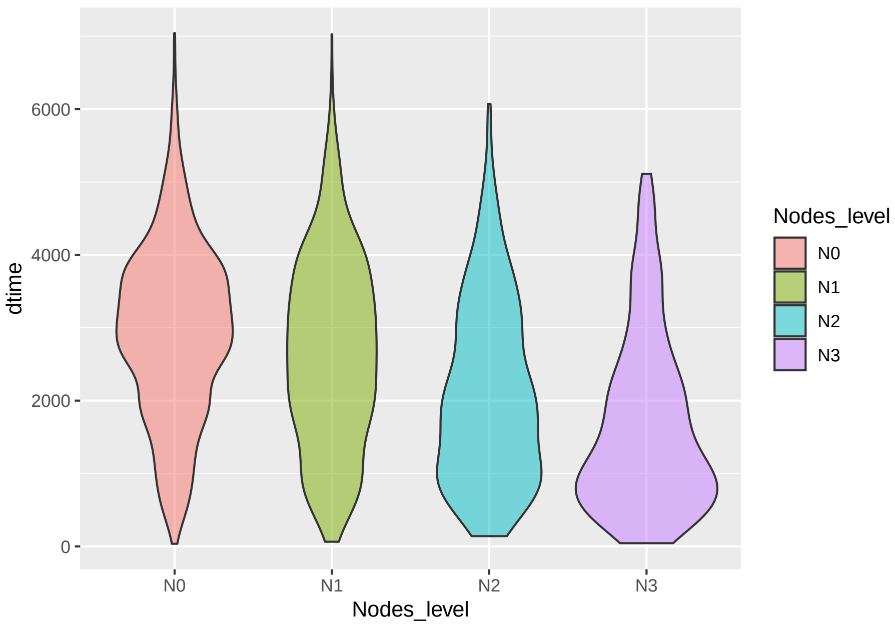
```
ggplot(data = rotterdam_recur, aes(x = size, y = drecurtime, fill = size)) +
  geom_violin(alpha = 0.5)
```



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.

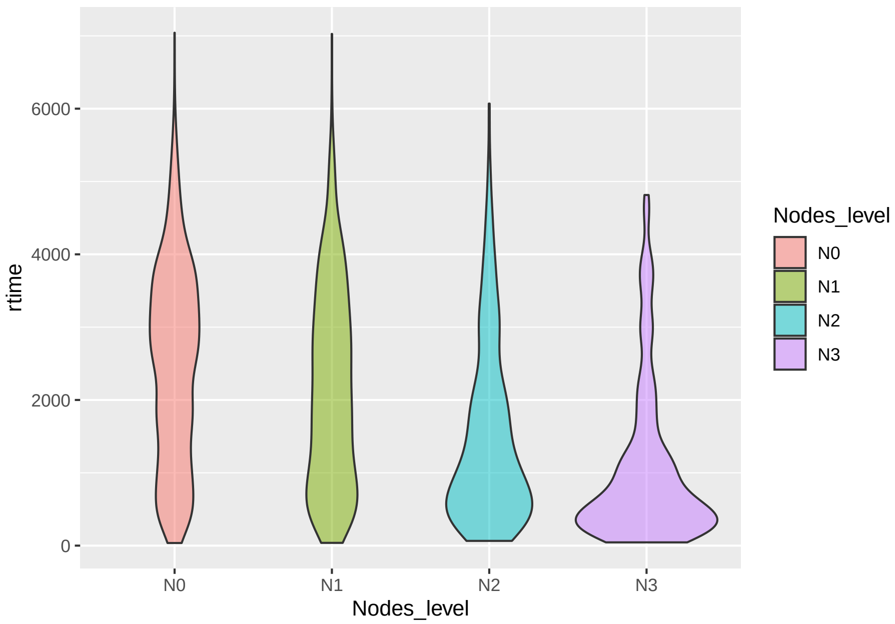
`Nodes_level` vs. `dtime`

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



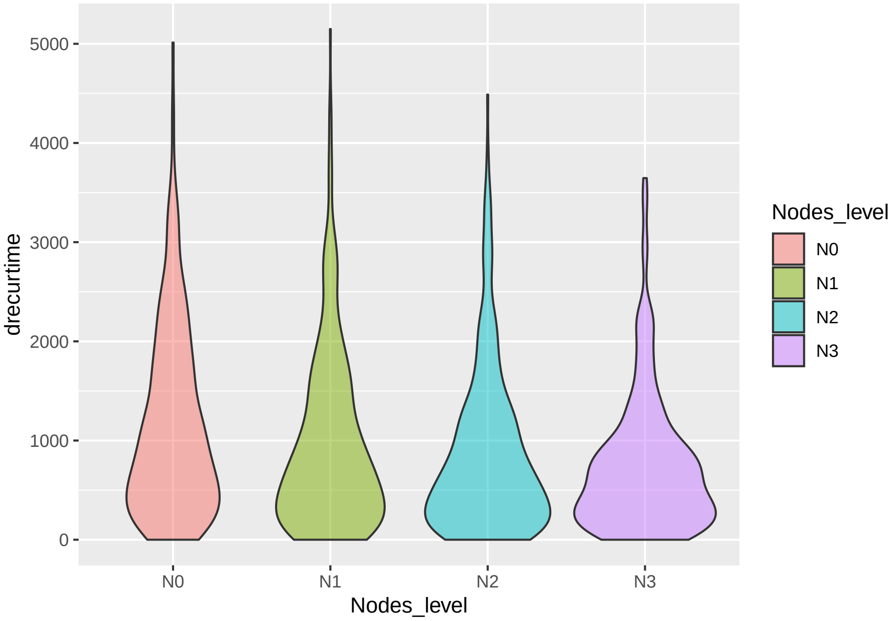
`Nodes_level` vs. `rtime`

```
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)
```



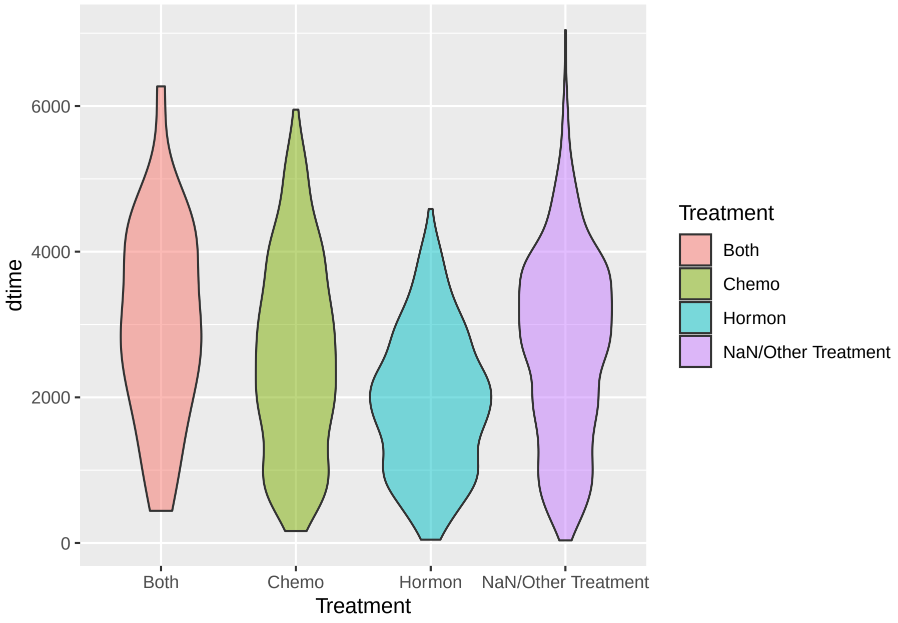
Similarly, `nodes` is also a factor impacting the life of breast cancer patients. In fact, for patients with high `Nodes_level`, it is typically considered they are either having metastasis of the cancer or already experiencing a regional recurrence of the cancer. Thus, we could see that most patients 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.

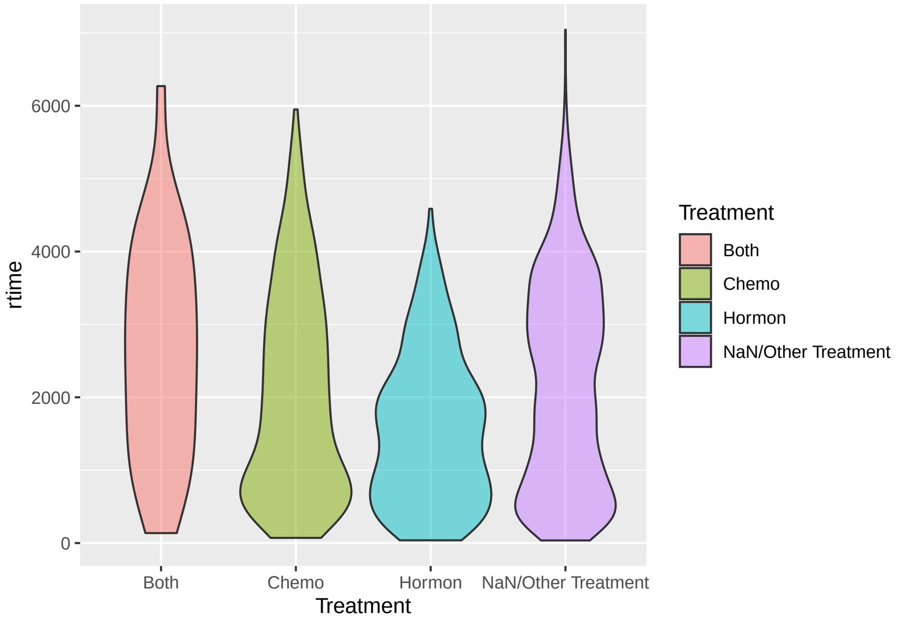
`Treatment` vs. `dtime`

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



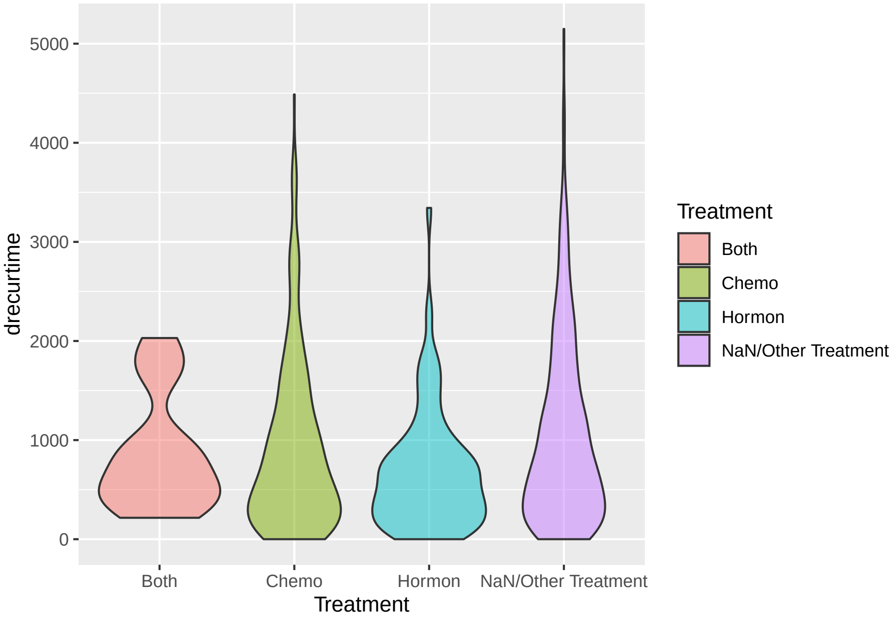
`Treatment` vs. `rtime`

```r
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 examing 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.

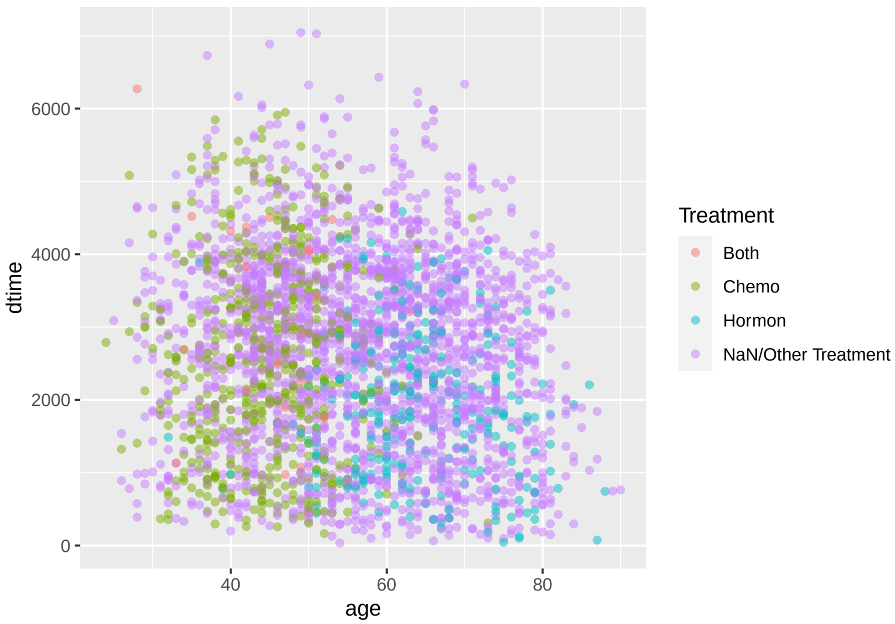
We also found a bi-model shape in `NaN/Other Traetment` group in `rtime` vs. `Treatment`. This may because the two peaks correponds to no treatment and other treatment separately, but currently we don't have more information investigating the true reason.

2.3.3 `age` + `Treatment` vs. Survival Times

`age` + `Treatment` vs. `dtime`

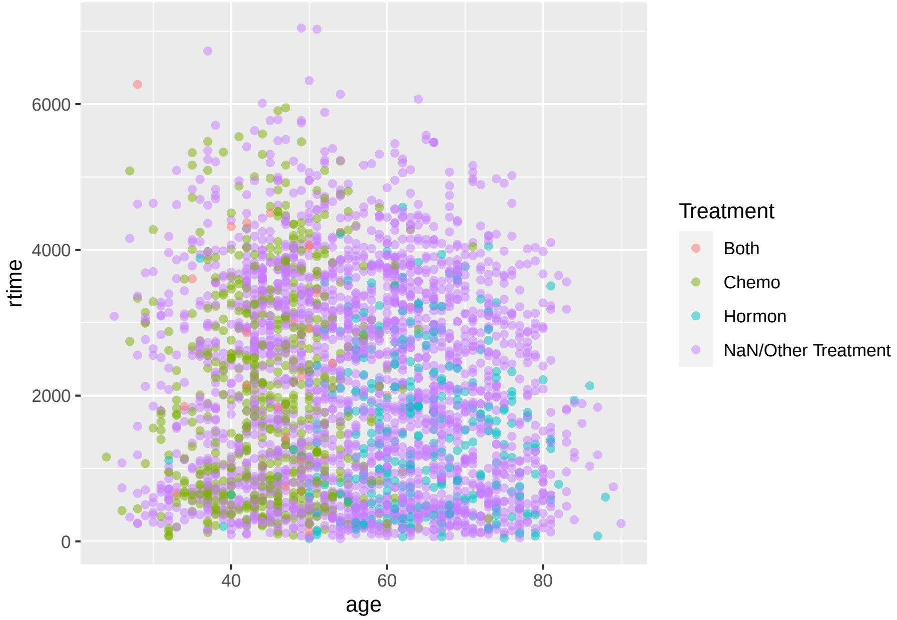
As we were visualizing for `Treatment` vs. `dtime`, we found that Hormontherapy generally has a weaker effect than Chemotherapy, but we think there might be some confounding variables that leads to such conclusion. One that we discovered is `age`:

```
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)
```

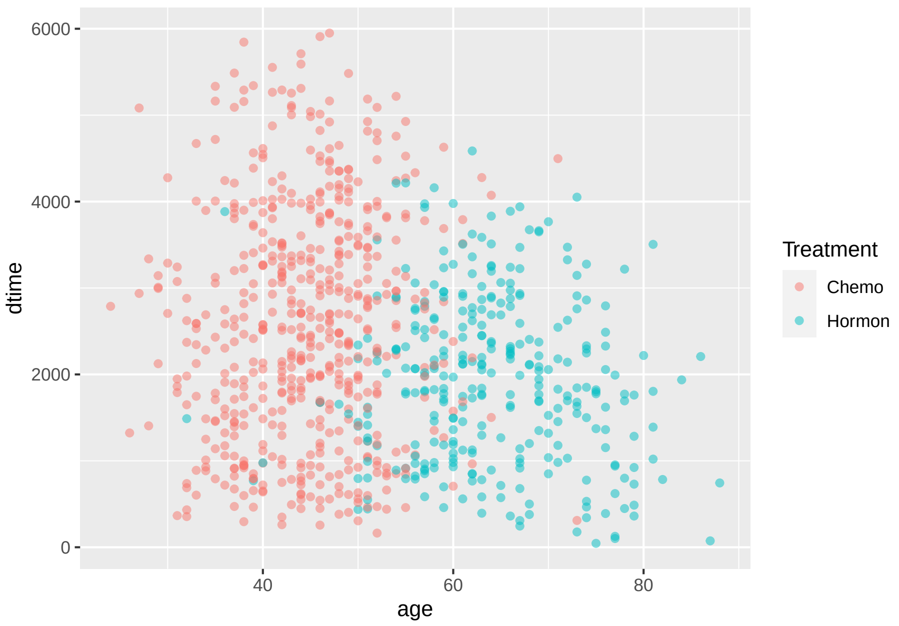


It might be difficult to see from the plots right now, so we decided to make a partial plot of the full plot by filtering the patients who did not take either treatment out.

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

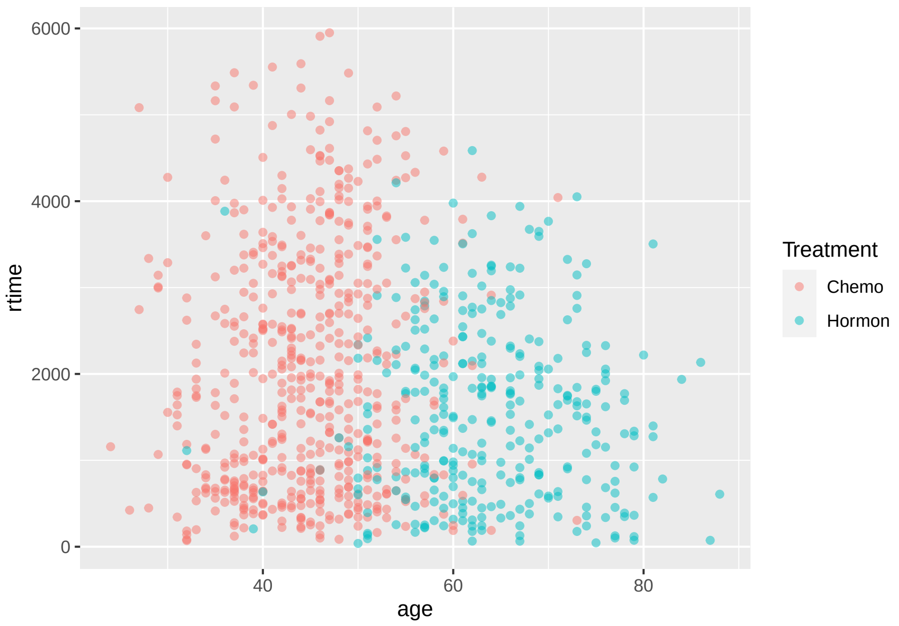
age + Treatment vs. dtime

```
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)
```

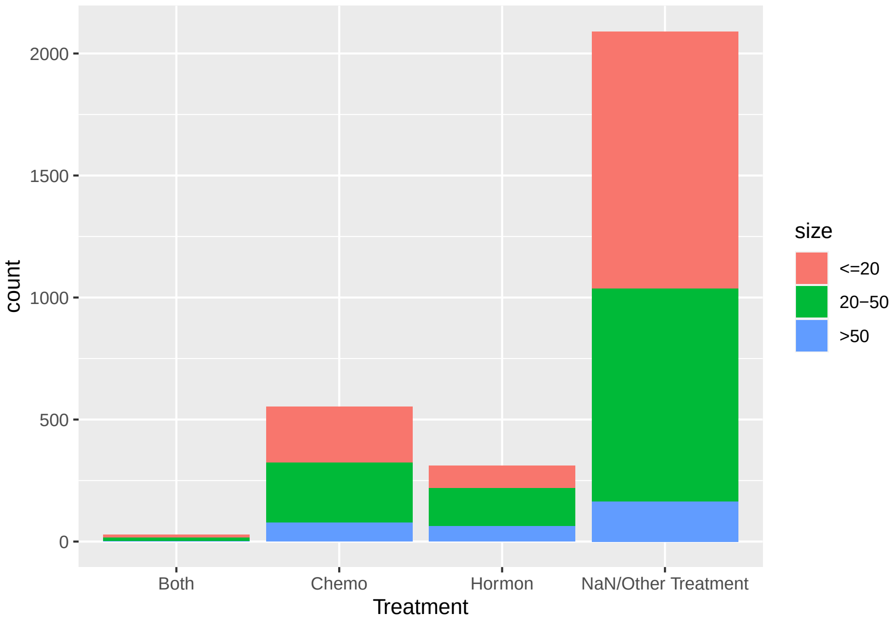


From the two plots above we could see that the patients who took either chemo therapy or hormon therapy are clearly clusterd. For the group who only took chemotherapy, most patients' age are located below 50 years old. For the group who only took hormontherapy, most patients' age are located above 50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

2.3.4 Treatment vs. TNM

Treatment vs. size

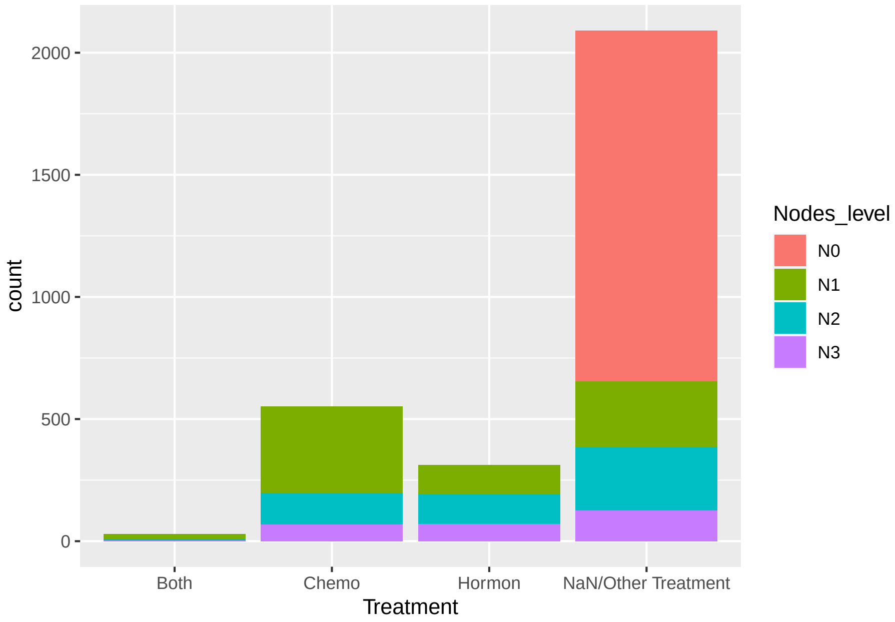
```
ggplot(data = rotterdam, aes(x = Treatment, fill = size)) +
  geom_bar(position = "stack")
```



We are also interested to see if there are any other factors that would affect a patient's decision on the treatment he/she takes other than his/her age. We thought we could stick with the TNM system and we started with tumor size. Generally, there is no distinction in the treatment taken among different sizes of tumor as we can see from the plot above.

Treatment vs. Nodes_level

```
ggplot(data = rotterdam, aes(x = Treatment, fill = Nodes_level)) +
  geom_bar(position = "stack")
```



This time we are checking if there are difference in `Treatment` with respect to different positive lymph node levels. Now we are onto something interesting. We can see that none of the `N0` patients in our dataset have taken either chemotherapy or hormontherapy. we think that it is possible that chemotherapy and hormontherapy are for severer patients and they might just be "overkill" for mild patients.

2.3.5 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 caner(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

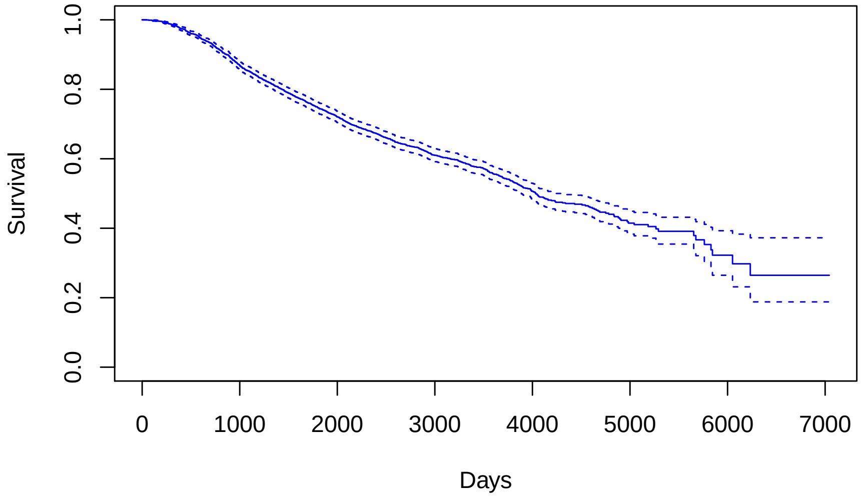
3.1 Loading Data

```r
data(rotterdam)
```

3.2 Kaplan-Miere estimator of the entire dataset

Death Time

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



```
mean(rotterdam$dtime)
```

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

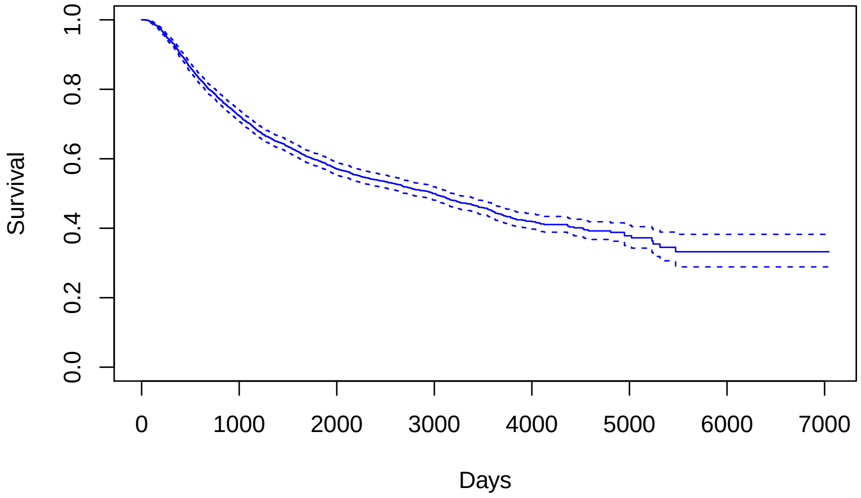
```
median(rotterdam$dtime)
```

```
## [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")
```



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

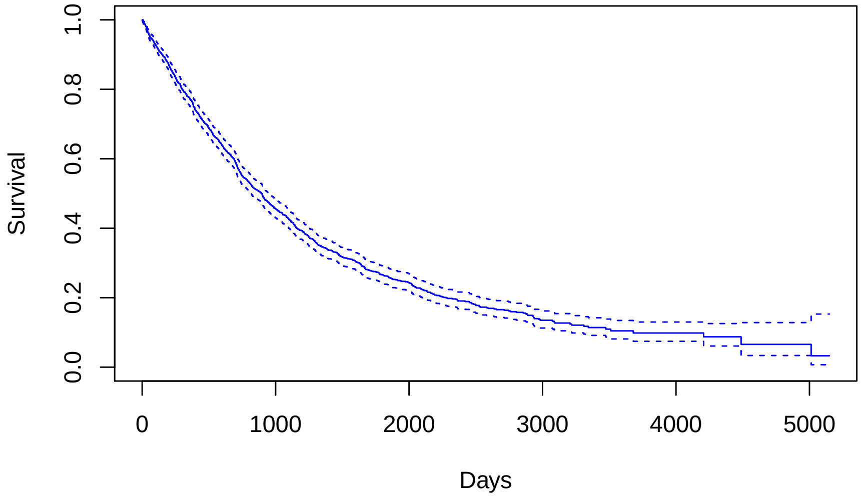
```
## [1] 2097.903
```

```r
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.

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



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

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

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

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

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

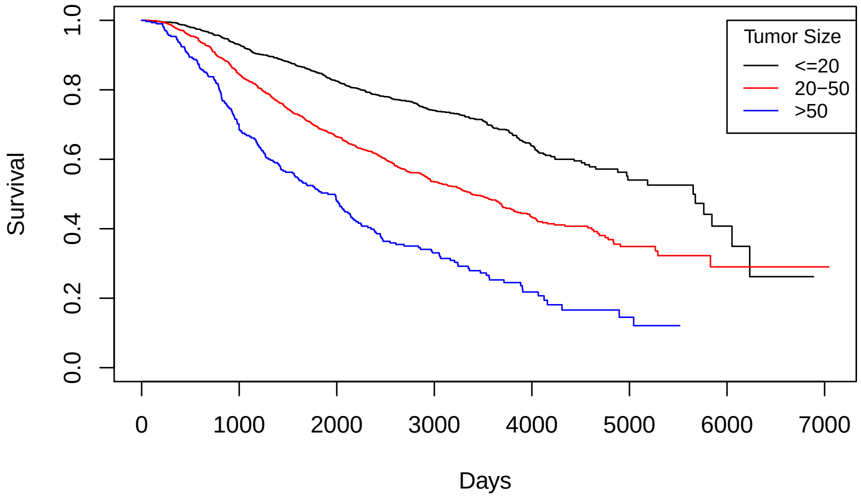
3.3 Kaplan-Miere estimator on different variables in rotterdam

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-Miere 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 categories of the variables.

3.3.1 size vs. Survival Times

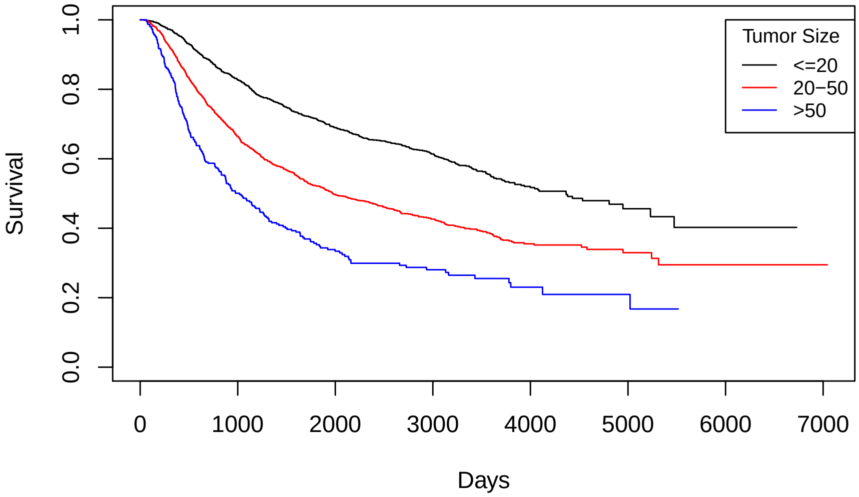
`size` vs. `dtime`

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



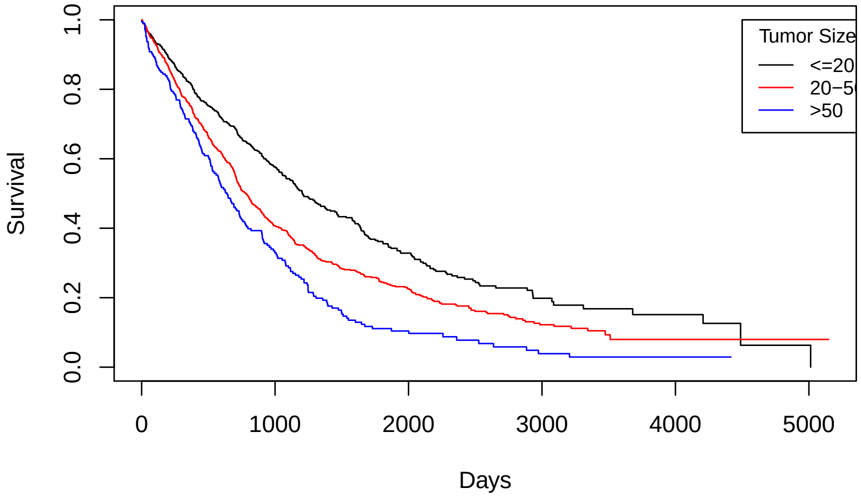
`size` vs. `rtime`

```
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("black", "red", "blue"), lty=1, cex=0.8,
       title="Tumor Size", text.font=6)
```



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="Survi
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)
```

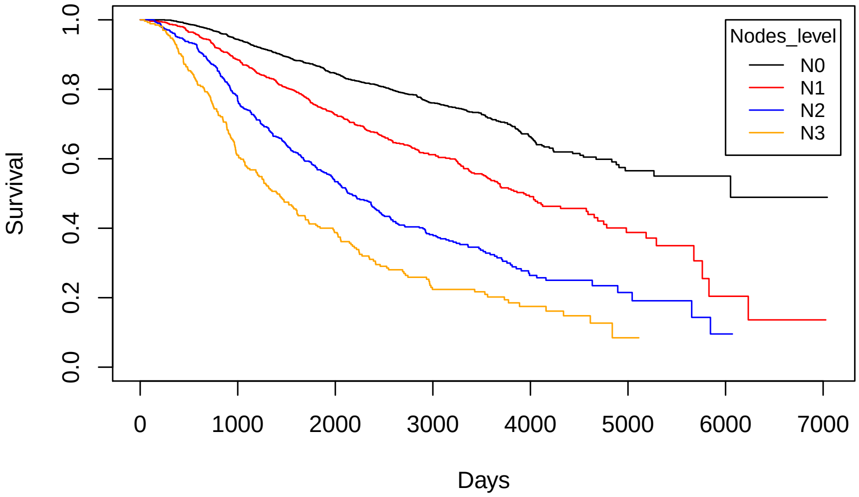


The trend is still the same as patients with smaller tumor size enjoy longer survival of death after recurrence, but the survival time now decreases much faster for all groups.

3.3.2 Nodes_level vs. Survival Times

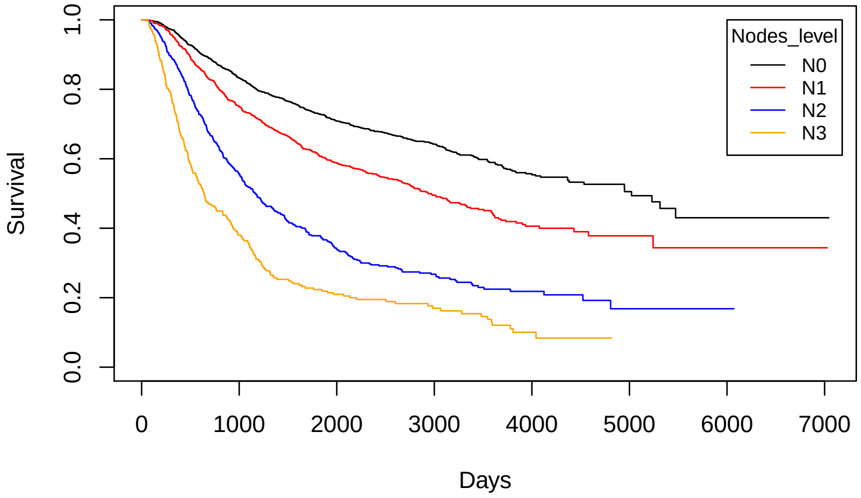
Nodes_level vs. dtime

```
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=
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)
```



`Nodes_level` vs. `rtime`

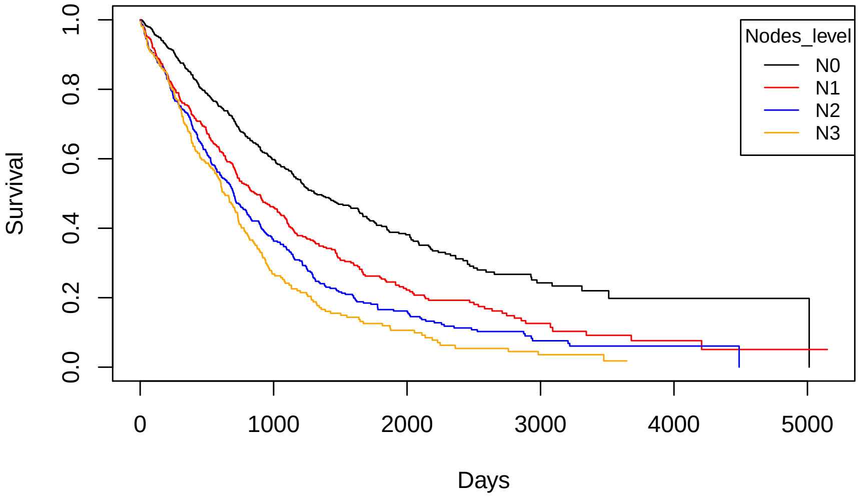
```r
KM_None_Recur <- survfit(Surv(rtime, recur) ~ Nodes_level, data = rotterdam)
plot(KM_None_Recur, conf.type = "plain", col = c("black","red","blue","orange"), xlab="Days", yla
legend(6000, 1, legend=c("N0", "N1","N2", "N3"),
       col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,
       title="Nodes_level", 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=
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)
```

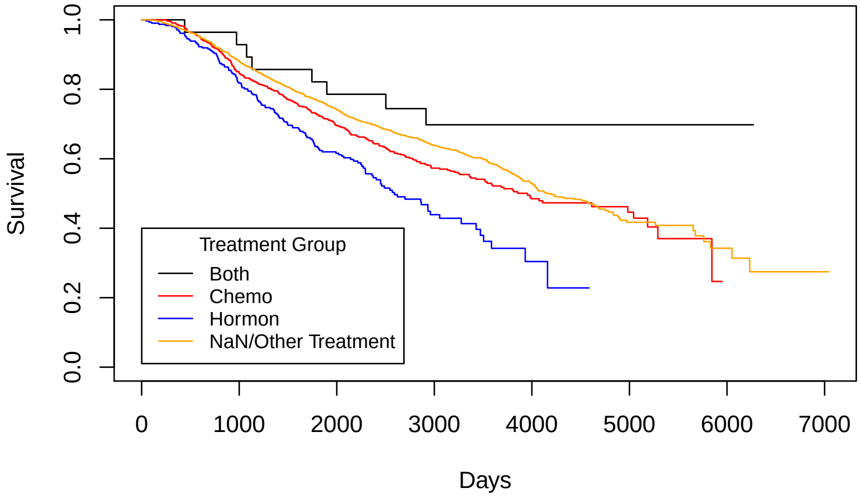


The trend is still the same as patients with fewer nodes tested positive enjoy longer survival of death after recurrence, but the survival time now decreases much faster for all groups, and the difference is small in groups `N1`, `N2` and `N3`.

3.3.3 Treatment vs. Survival Times

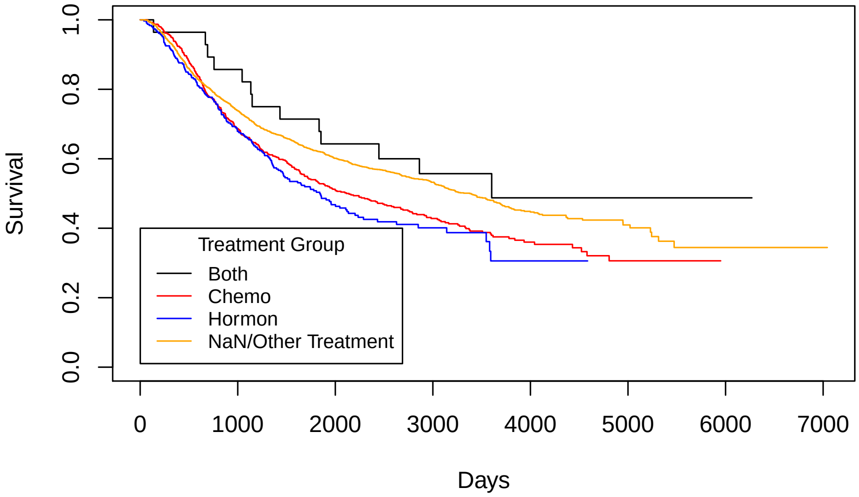
`Treatment` vs. `dtime`

```r
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`

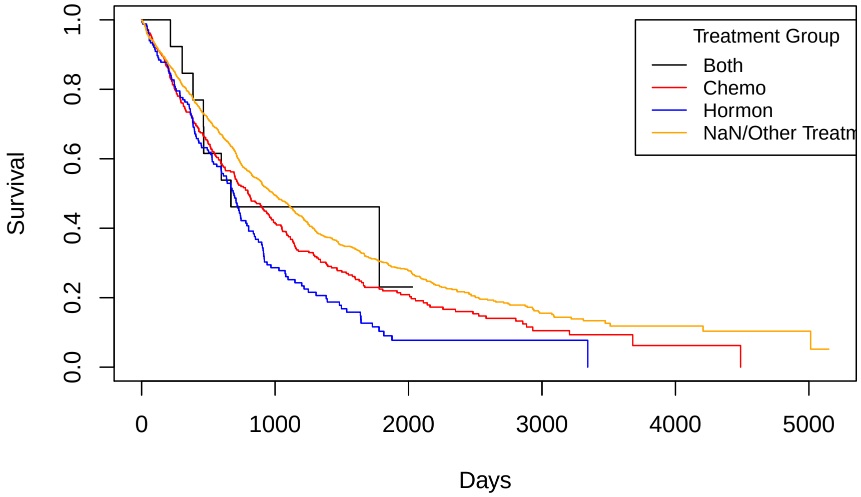
```
KM_Treatment_Recur <- survfit(Surv(rtime, recur) ~ Treatment, data = rotterdam)
plot(KM_Treatment_Recur, conf.int = FALSE, col = c("black", "red", "blue", "orange"),
legend(1, 0.4, legend=c("Both", "Chemo","Hormon", "NaN/Other Treatment"),
       col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,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



As we have discussed in Chapter 2, we know that generally chemotherapy is used on patients with age lower than 50 years old and hormontherapy 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 hormontherapy 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

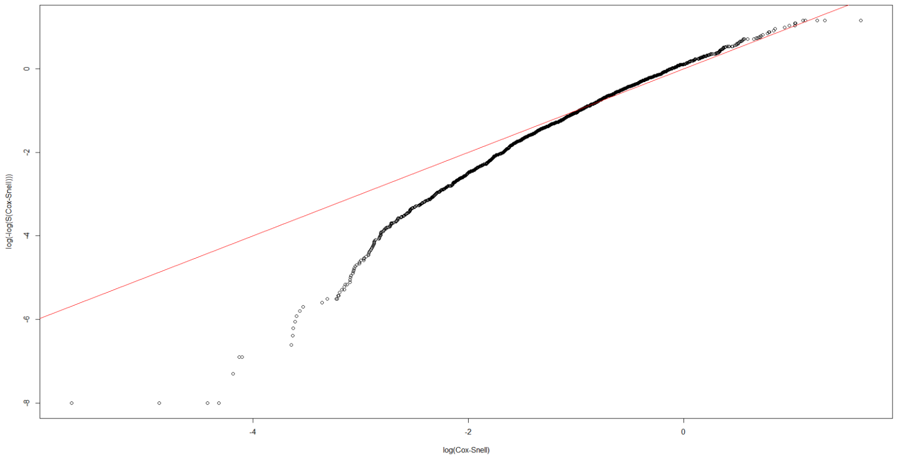
```
# The Cox-Snell function takes as inputs
# 1. A vector of Cox-Snell residuals created by the user based on the model being evaluated,
# 2. A status vector
# 3. Optional x- and y- limits for the resulting plot

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.

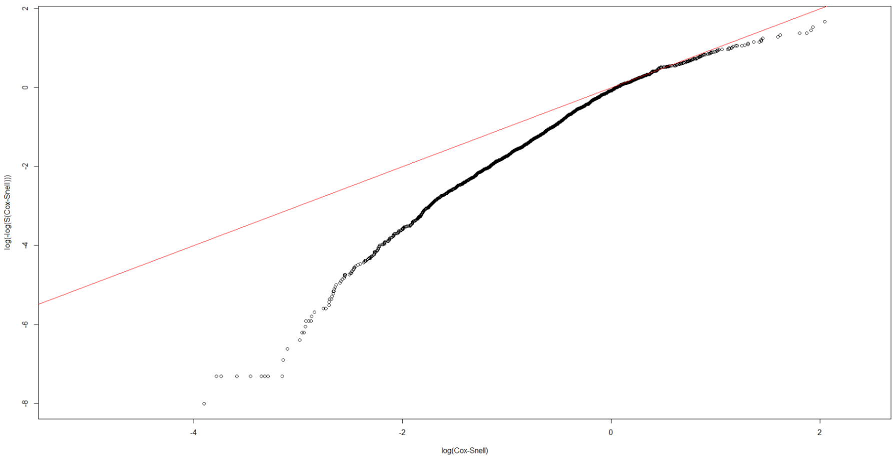
```
Dexp <- survreg(Surv(dtime, death) ~ Treatment + size + nodes + age, dist='exponential', data=rot
Dexp
```

```
## 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
```



```
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
```



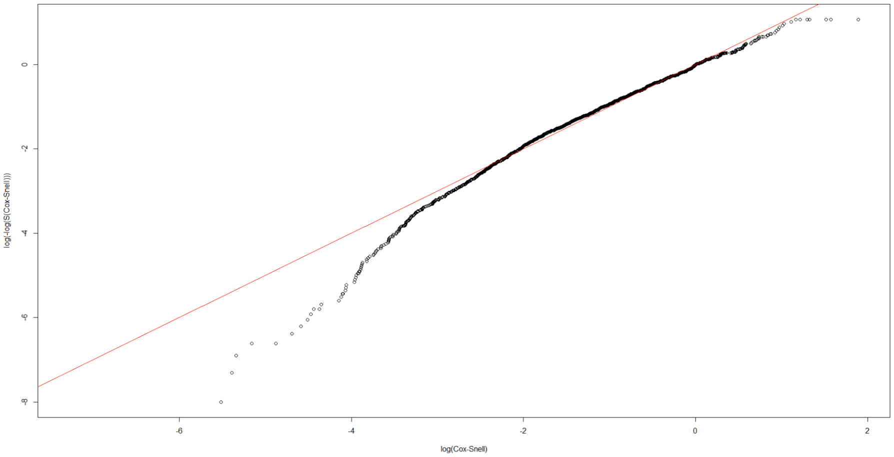
The Exponential does not seem adequate in this case.

3.4.3 Weibull models

Next we will examine the adequacy of Weibull model.

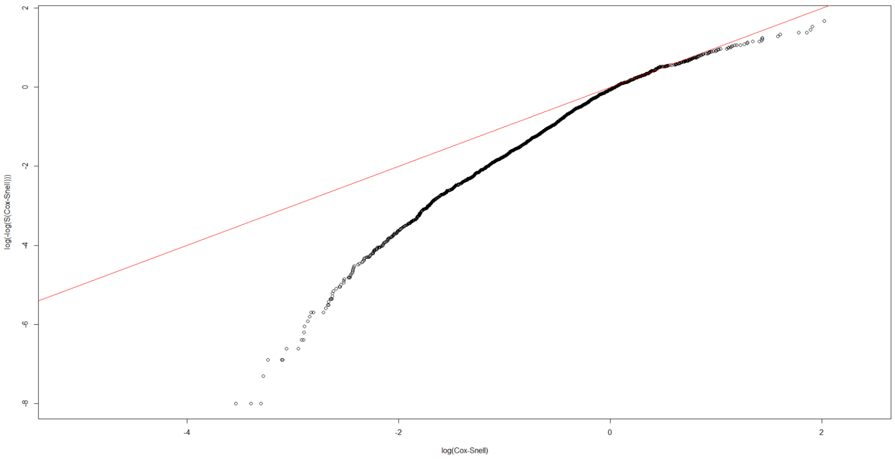
```
Dweibull <- survreg(Surv(dtime, death) ~ Treatment + size + nodes + age, dist='weibull', data=rot
Dweibull
```

```
## 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
```



```
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.

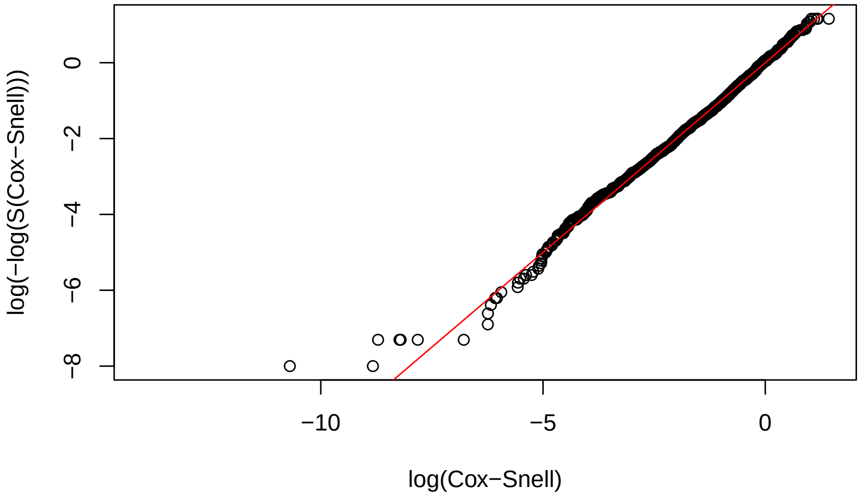
3.4.4 Log-normal models

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

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

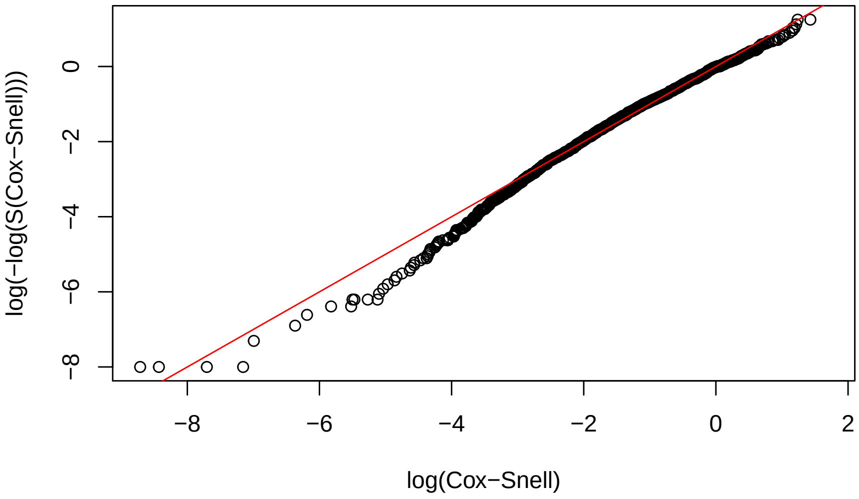
```
CS_LnormD <- -log(1 - plnorm(rotterdam$dtime, 9.709442268-0.431016329*(rotterdam$Treatm
                                                   -0.346351742*(rotterdam$Treatment=="H
                                                   -0.423626557*(rotterdam$Treatment=="N
                                                   -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_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 `dtime` and `rtime` vs. `Treatment` + `size` + `nodes` + `age`.

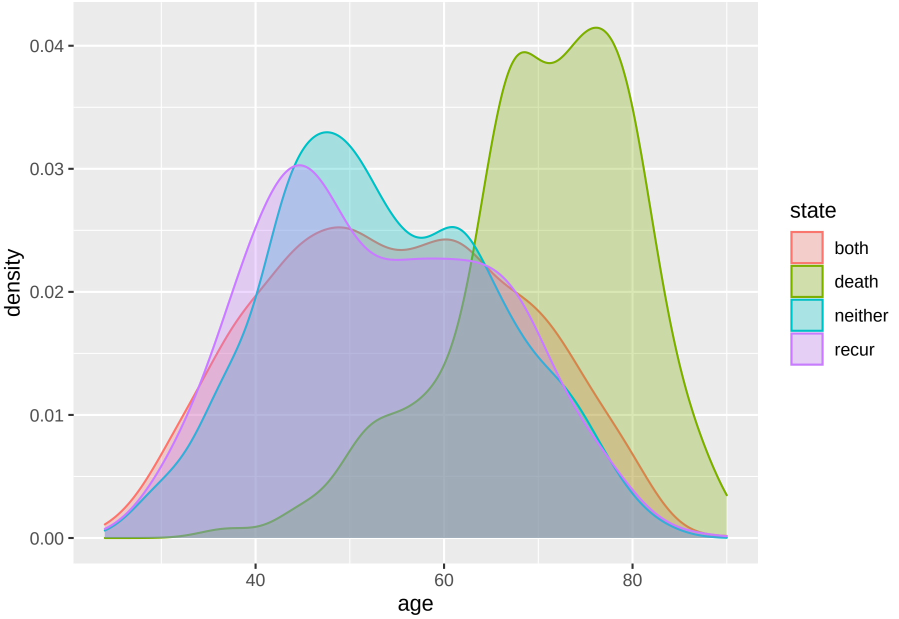
As to why the Log-normal model would be suitable for the data, we are not sure. One thing to consider is that the non-monotonic implication of the data casted by the log-normal model.

3.4.5 Positive coefficient of `age`

In the `rtime ~ Treatment + size + nodes + age` models, it is surprised to see that the coefficient of `age` is positive, though the value is small. One possible explanation could be the idea of “competing events” and “competing risk”.

In this scenario, having recurred breast cancer and being dead could be somewhat “competing events”. Though they are not completely “cannot happen on one person at the same time”, it is still reasonable to think that for older patients, it is more likely to die from breast cancer or other complications than being cancer-free for years and then having breast cancer recurred; whereas for younger people, the risk of having recurrent breast cancer could be higher than being dead from the first breast cancer.

```
rotterdam_new <- rotterdam %>%
  mutate(state = ifelse(death == 1 & recur == 0, "death",
                ifelse(death == 0 & recur == 1, "recur",
                ifelse(death == 1 & recur == 1, "both", "neither"))))

ggplot(rotterdam_new, aes(x=age, color=state, fill=state)) +
  geom_density(alpha=0.3)
```



The plot above also helps verifying the idea. We can find that the group who has been dead during the study but has never had breast cancer recurred (the green one) tends to be older than others.

The positive coefficient of `age` in the log-normal model may seem indicating a protective effect of being old against having breast cancer recurred, but this should not be the true case. It is very likely that this is caused by the competing risk between being dead and having breast cancer recurred for people in different age group. The elderly patients are less likely to suffer from recurrent breast cancer because they are more likely to die from breast cancer or other complications during the treatment after the first diagnosis.

3.5 Cox-PH model:

Another very important part of a survival analysis is building and interpreting the coxph model. With this model, once we have proof of the validation of the PH assumption of the model, we could have nice coefficients to interpret as the logarithm of HR(hazard ratio) among different groups. We will be following the way that we built our parametric models, by using Survival times vs. `age` + `size` + `nodes` + `Treatment`.

3.5.1 `dtime` coxph 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
```

First we should check if the PH assumption holds. We will be using a formal test.

```
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
```

One of the variable, `age`, has p-value smaller than 0.05, which indicates the violation of the PH assumption. So we can not intepret the coefficients right away.

One way we came up to solve this problem is to put stratification on `age` by putting a `strata()` on `age` when fitting the model. What this does is to recognize the correlation between `age` and the `dtime`, but not actually including it in our model results.

```
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
```

Now we can see that all the variables' p-value are greater than 0.05 which indicates the PH assumption holds in our model. Now we can finally intepret the coefficients.

For treatments, we can see that Chemotherapy, Hormontherapy and Other Treatment all have higher risks than patients receiving both therapies. However, note that all three p-values are greater than 0.05, which means that we do not have significant evidence for such relationships. We think that this is because there are too few data point receiving both therapies.

Now let's look at tumor size. It is clear that patients with larger tumor size enjoy higher risks, and as the size grows, the risk is also increasing. such relationship is significant as we can see from the p-values since they are all less than 0.05.

For lymph nodes tested positive, we could see that for 1 unit increase in the number of nodes, the risk will increase by a multiplicative of 1.08. This relationship is also significant as the p-value is smaller than 0.05.

```
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) ~ strata(Treatment) + size + nodes + age, data=rotte
m_recur
```

```
## Call:
## coxph(formula = Surv(rtime, recur) ~ strata(Treatment) + size + 
##     nodes + age, data = rotterdam)
## 
##                 coef exp(coef)  se(coef)      z        p
## size20-50   0.396037  1.485924  0.057899  6.840 7.91e-12
## size>50     0.693201  2.000107  0.087565  7.916 2.44e-15
## nodes       0.080220  1.083525  0.004576 17.529  < 2e-16
## age        -0.008805  0.991233  0.002315 -3.804 0.000142
## 
## Likelihood ratio test=404.3  on 4 df, p=< 2.2e-16
## n= 2982, number of events= 1518
```

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

```
##          chisq df       p
## size   31.3790  2 1.5e-07
## nodes   9.8790  1  0.0017
## age     0.0979  1  0.7544
## GLOBAL 35.2423  4 4.1e-07
```

Chapter 4

Conclusion

From our research, we have seen and proven that though treatment(in our dataset) might appear to have different effect on patients, by including confounder and other factors, we do not have enough evidence that the difference is solid. For the TNM diagnostics of breast cancer, though we do not have data about metastasis, we still had significant result for the relationship between tumor size and number of lymph nodes tested positive. We have also seen that the most formidable aspect of breast cancer(or maybe all cancer) is the recurrence, as in our research there's not much time left if a patients tumor have recurred. But generally, breast cancer has a relative long survival and patients diagnosed with breast cancer should not be pessimistic but rather face reality and seek proper treatment. In this way, one might achieve as long survival as possible.

Of course, there are many limitations within our research and much more that we could explore in the realm of breast cancer. We will list a few for future study below.

4.1 Limitation

- As we have mentioned in Motivation, the TNM system for diagnosis of breast cancer relies on three important indicators, **Tumor**, **Node**, **Metastasis**. Now in our research, the dataset already contains information about tumor size and number of lymph nodes tested positive, but it does not contain any information about metastasis. However, as we have mentioned in Chapter 2, these information would not exist until a patient has developed a stage IV breast cancer. We don't know exactly if these patients were left out for a reason, but we still think that a more generalized subject pool would make the result of the research more convincing and more generalized. For future studies, we think that not only stage IV breast cancer patients but also stage I patients should all be recorded as part of the dataset.

- In Chapter 2, we compared the effect and the condition for choosing between chemotherapy and hormontherapy. We have had good results discerning the patients taking different therapy by age, as younger patients tend to take chemotherapy more and older patients tend to take hormontherapy more. And we have also seen that most patients with no lymph nodes tested positive choose neither of chemotherapy and hormontherapy. There might be other therapies that better suit mild condition patients. Additionally, nowadays, there are way more categories of therapies for breast cancer and many of them may have better effect on patients and less discerning characteristics like age with chemotherapy and hormontherapy. In future studies, these missing treatment shall be recorded, as it is critical to studying if the new methods are more effective and lasting treatment for breast cancer patients.

- In Chapter 3, we found out that in parametric models for recurrence vs. predictors, the coefficient for age is positive, which suggests that the older one patient is, the less likely he/she have recurred cancer. This counter-intuitive fact was explained by introducing the concept of competing risks. However, in the data that we obtained, there is no record of each patients' cause of death, and thus we could not conduct further analysis on what would be possible competing event that makes this relationship exist. In future study, if possible, researcher should try to record for patients with exact time of death their cause of death. In this manner, we might be able to study the causes of death for patients with breast cancer but did not die from the cancer itself.

4.2 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).