Rotterdam

JACK TAN, YIMING MIAO

2021-03-14

Contents

Chapter 1

Motivation

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

In addition, cancer cells are given a grade when they are removed from the breast and checked in the lab. Based on how much they look like normal cells, there are three grades of cancer cells:

Grade 1 / well differentiated: The cells are slower-growing, and look more like normal breast tissue, meaning that cancer is less likely to spread.

Grade 2 / moderately differentiated. The cells are growing at a speed of and look like cells somewhere between grades 1 and 3.

Grade 3 / poorly differentiated: The cells look very different from normal cells, meaning a faster-growing cancer that's more likely to spread.

We will use data related to information above 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 of 2982 breast cancer patients. 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` standing for the size of the tumor, `nodes` standing for how many lymph nodes are test cancer positive, and `grade` standing for an metric of metastasis, so we have all three criterions 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')
```

2.2.1.0.1 `size` for T(Tumor):

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

We can see that most of the patients in the database have tumor size smaller than 50mm.

2.2.1.0.2 `nodes` for N(Node):

The number of lymph nodes tested positive is another important measure in the TNM system.

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
```

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

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

2.2.1.0.3 grade for M(Metastasis):

```
## # 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 in grade II or grade III, which means that the cancer cells may start shifting to other parts of their body.

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.

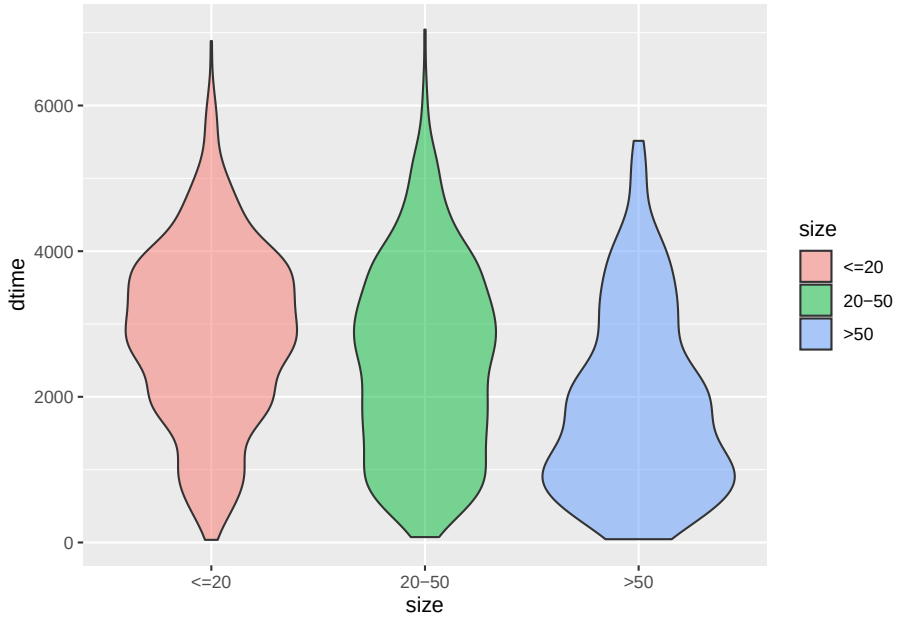
2.3 Data visualizations and exploration

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.

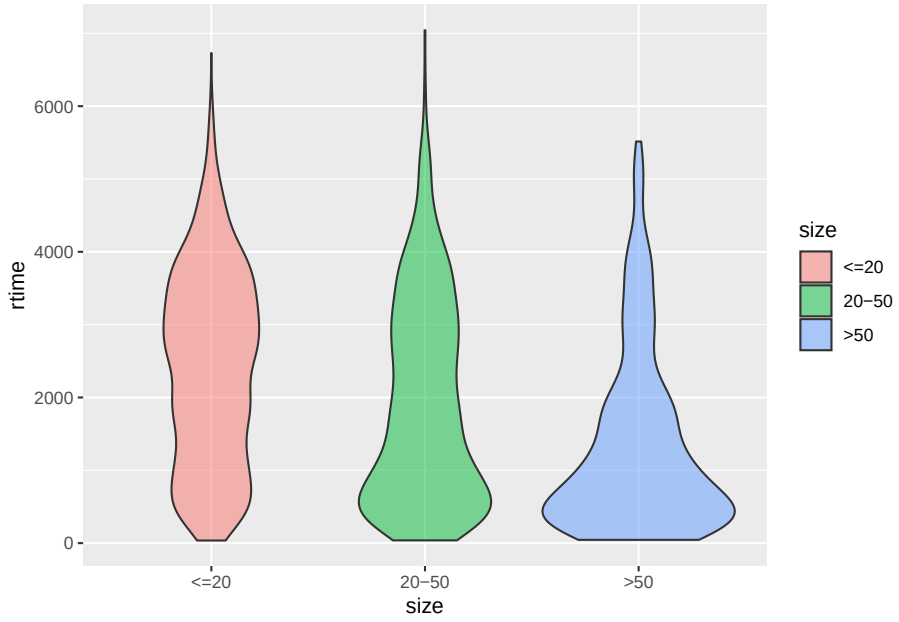
`size` vs. `dtime`

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



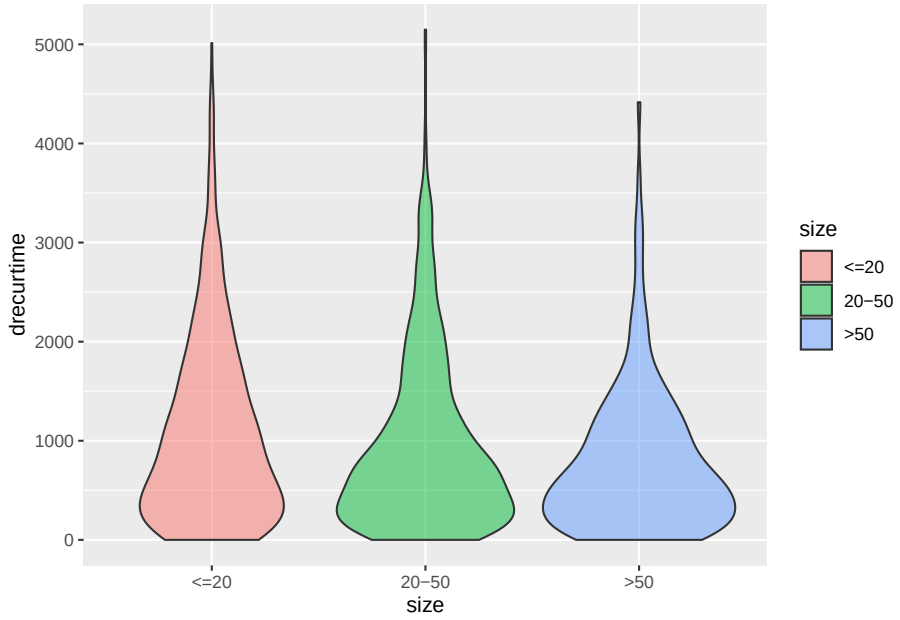
`size` vs. `rtime`

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



`size` vs. `drecurtime`

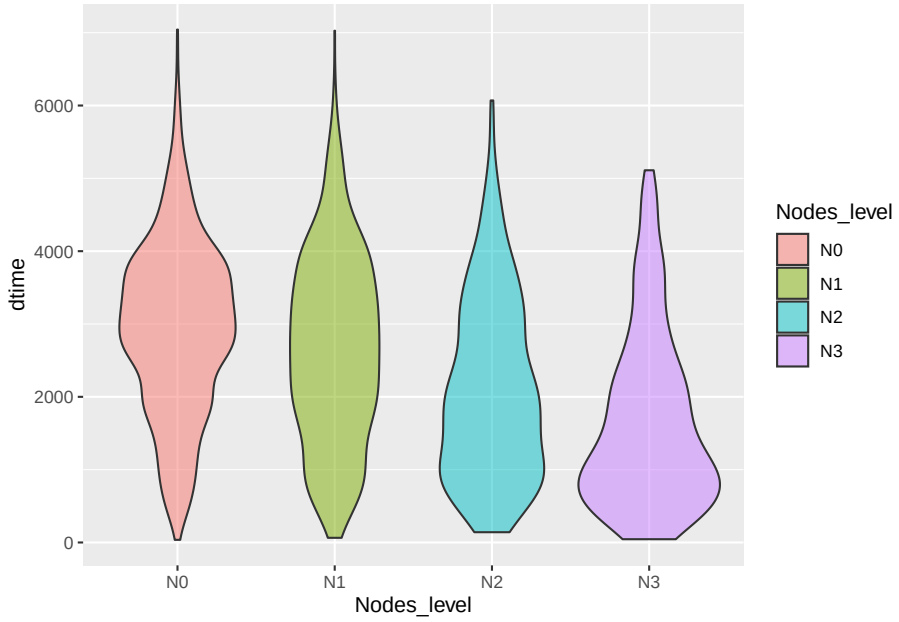
```
ggplot(data = rotterdam_recur, aes(x = size, y = drecurtime, fill = size)) +
  geom_violin(alpha = 0.5)
```



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.

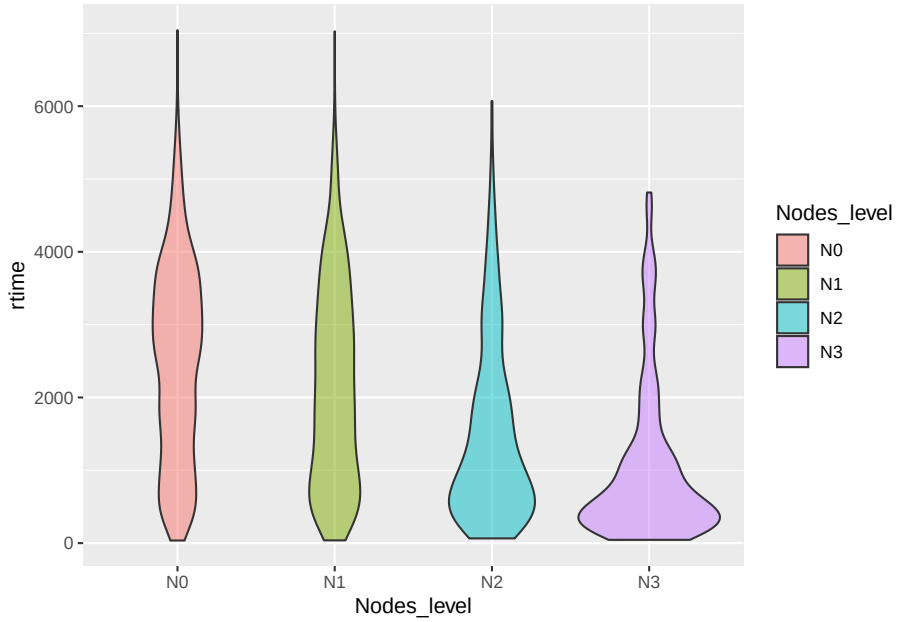
`Nodes_level` vs. `dtime`

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



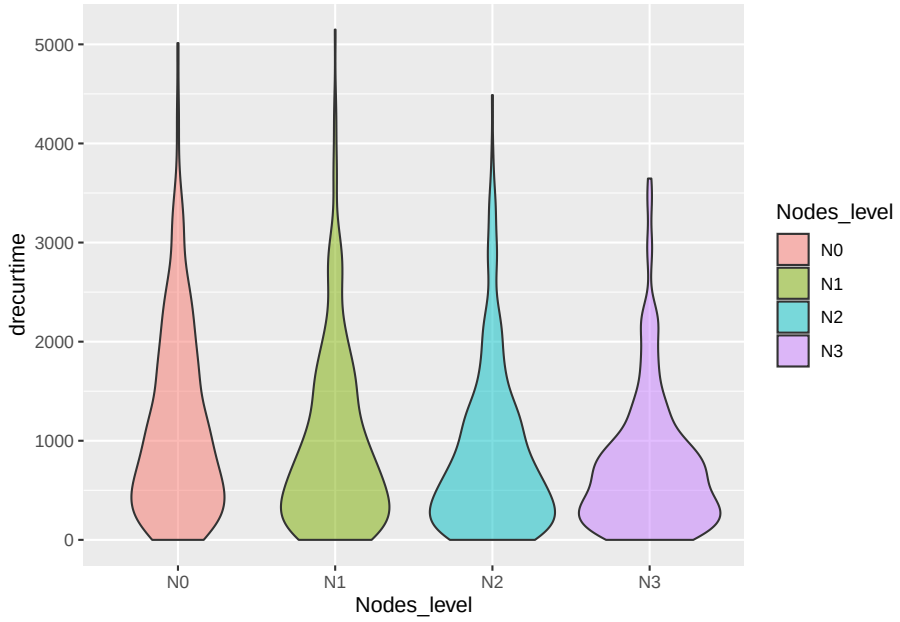
`Nodes_level` vs. `rtime`

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



`Nodes_level` vs. `drecurtime`

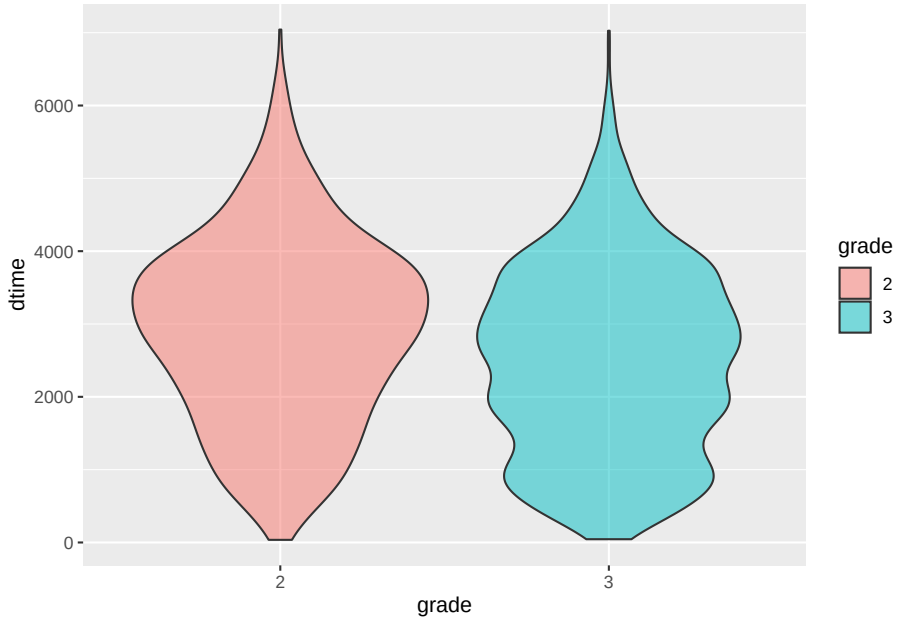
```
ggplot(data = rotterdam_recur, aes(x = Nodes_level, y = drecurtime, fill = Nodes_level)
  geom_violin(alpha = 0.5)
```



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.

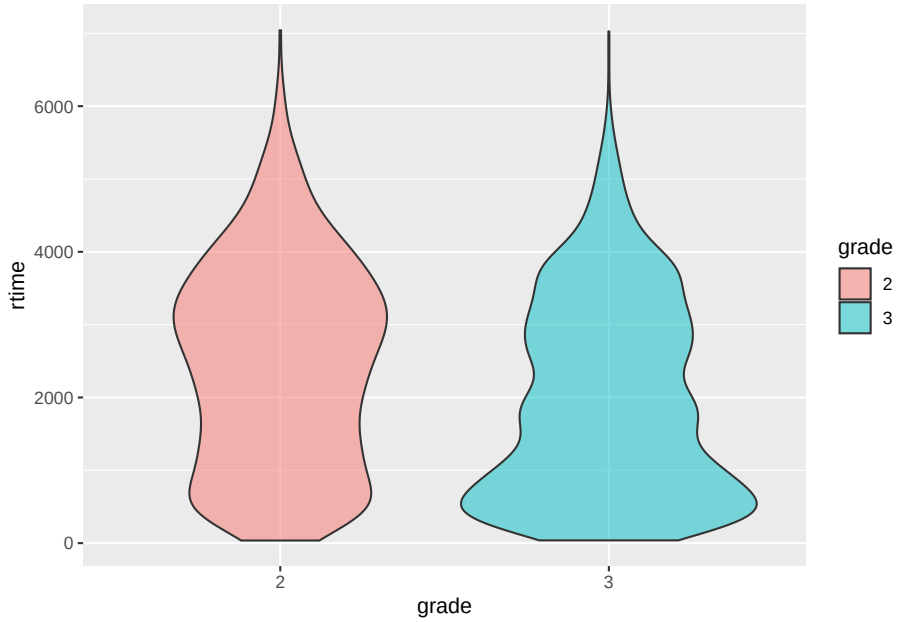
`grade` vs. `dtime`

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



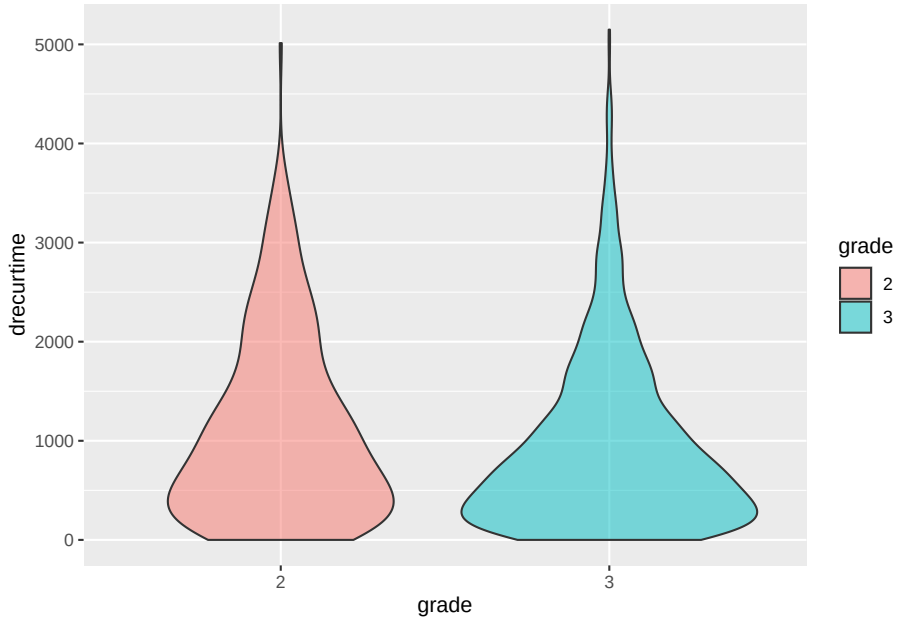
`grade` vs. `rtime`

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



`grade` vs. `drecurtime`

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



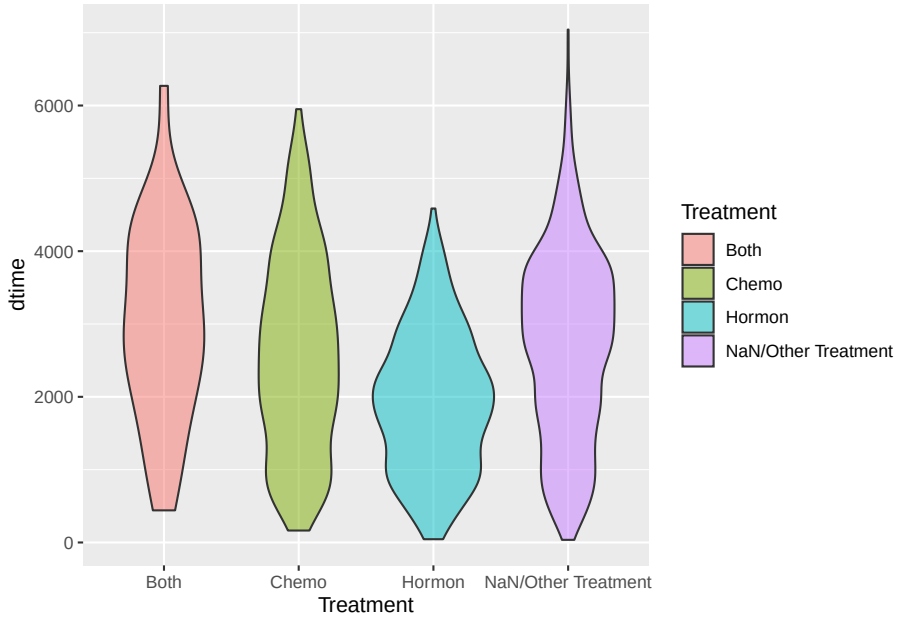
We noticed that a larger proportion of patients in grade 3 survived <2000 days than those in grade 2, and pretty much of patients in grade 3 suffer from recurrence of breast cancer in less than 1000 days.

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.

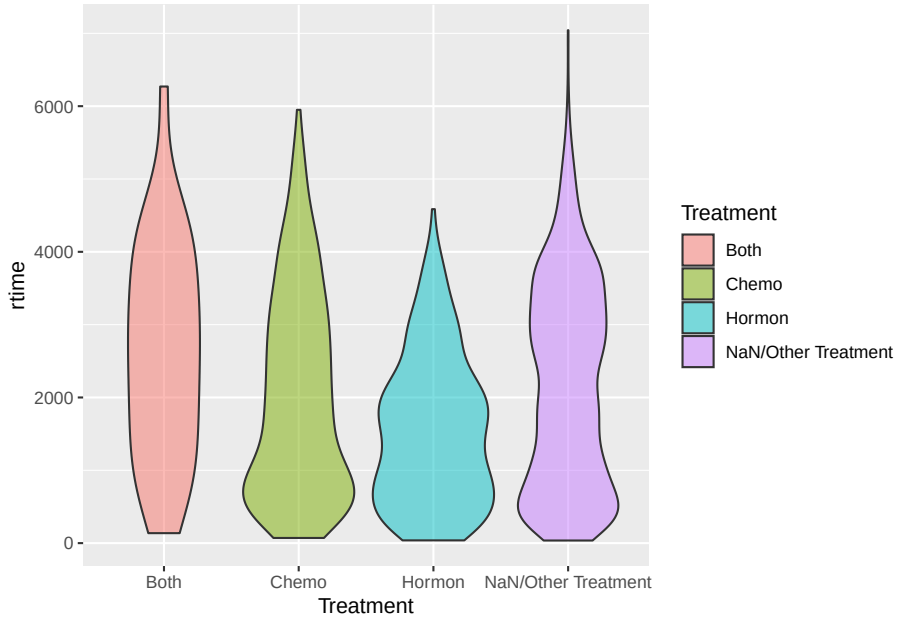
`Treatment` vs. `dtime`

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



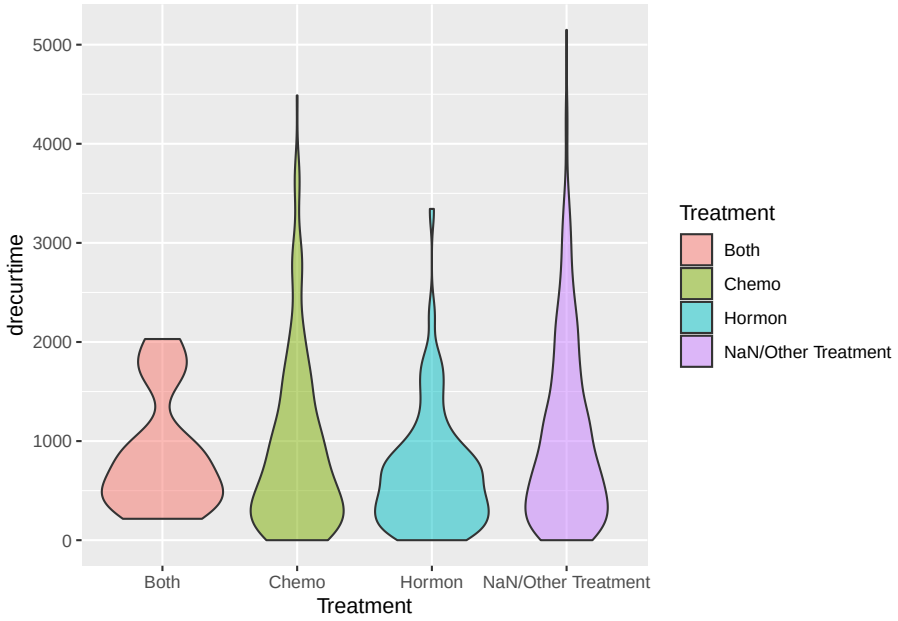
`Treatment` vs. `rtime`

```
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.

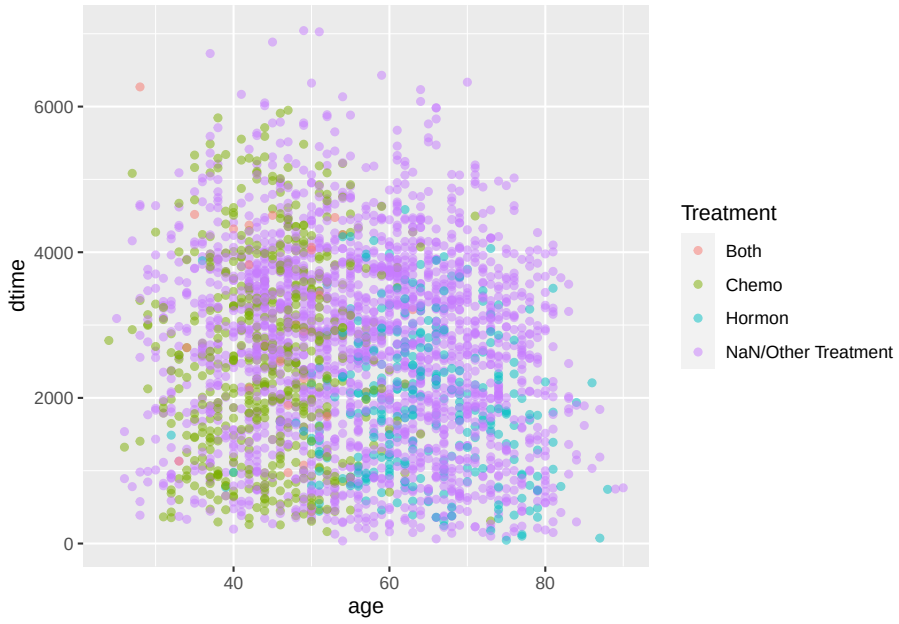
We also found a bi-model shape in `NaN/Other Traetment` group in `rtime` vs. `Treatment`. This may because the two peaks correponds to no treatment and other treatment separately, but currently we don't have more information investigating the true reason.

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

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

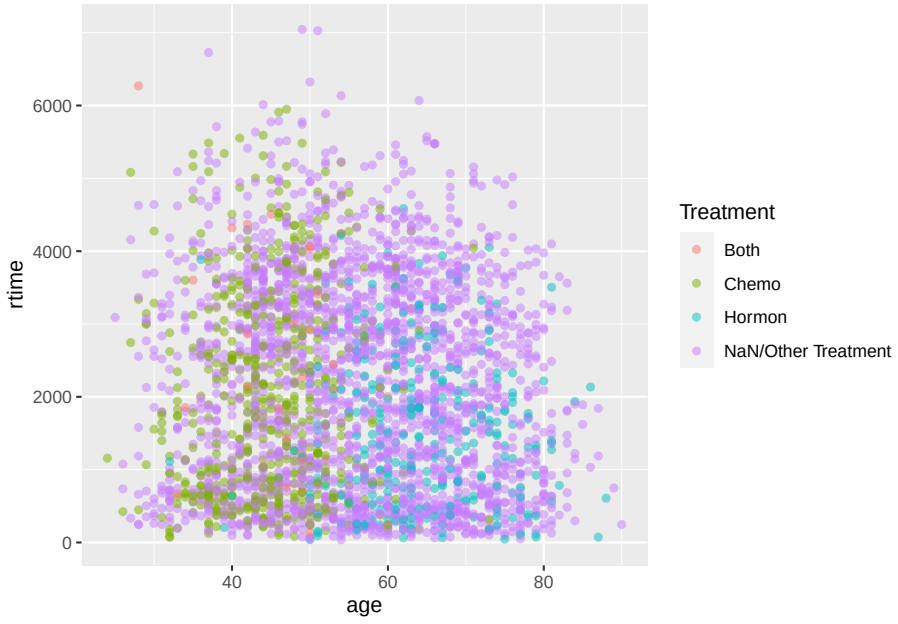
As we were visualizing for `Treatment` vs. `dtime`, we found that Hormontherapy generally has a weaker effect than Chemotherapy, but we think there might be some confounding variables that leads to such conclusion. One that we discovered is `age`:

```
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)
```

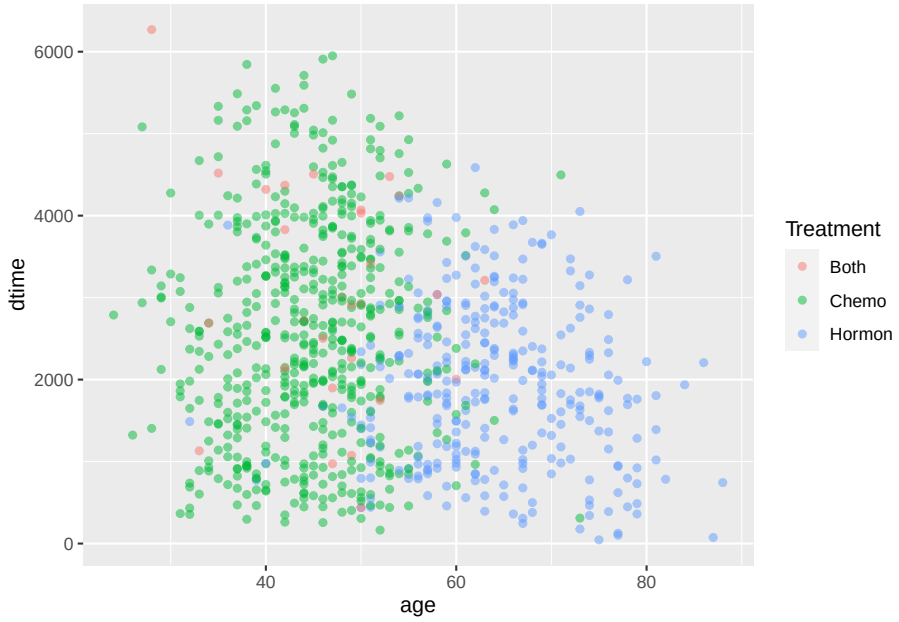


It might be difficult to see from the plots right now, so we decided to make a partial plot of the full plot by filtering the patients who did not take either treatment out.

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

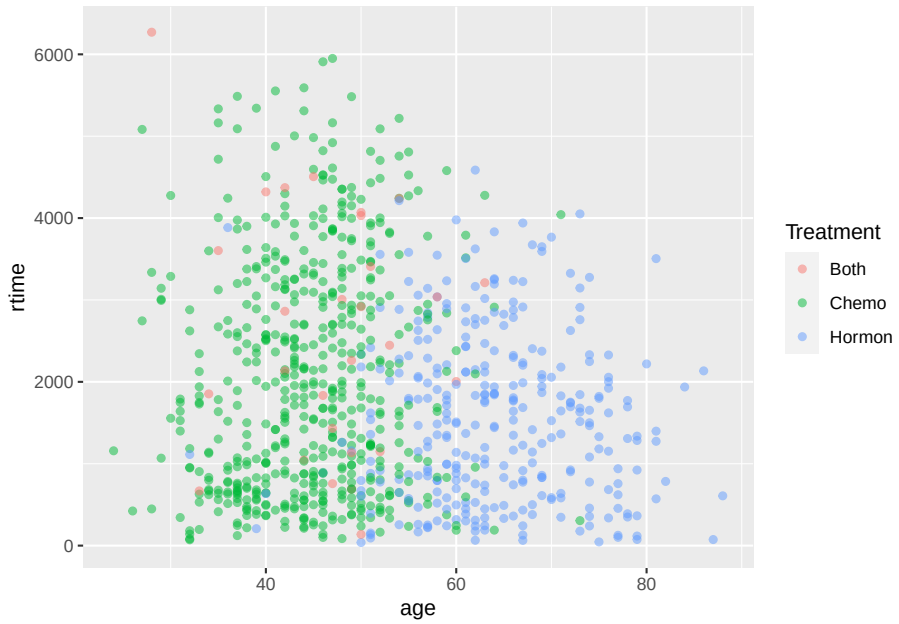
`age` + `Treatment` vs. `dtime`

```
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)
```

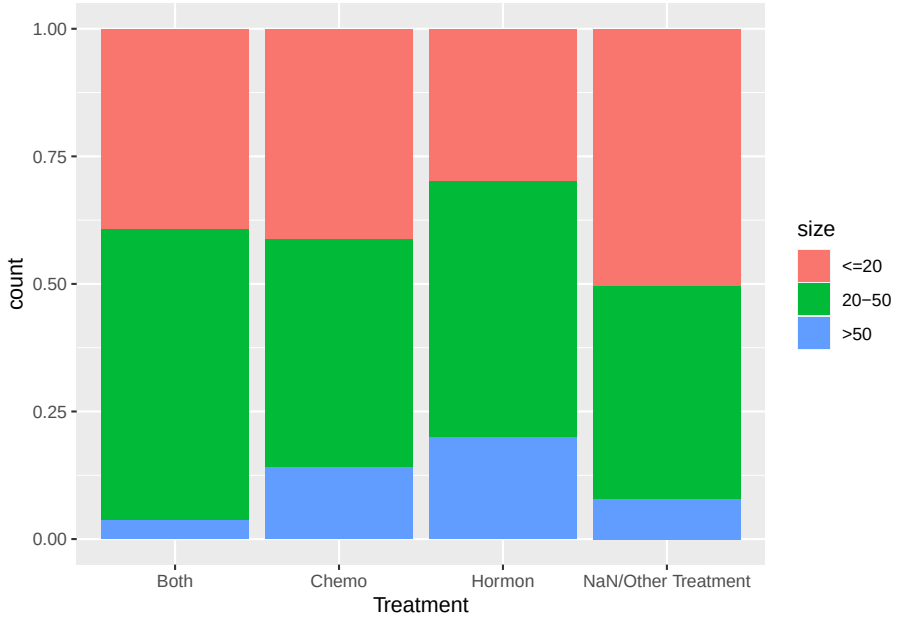


From the two plots above we could see that the patients who took either chemo therapy or hormon therapy are clearly clusterd. For the group who only took chemotherapy, most patients' `age` are located below 50 years old. For the group who only took hormontherapy, most patients' age are located above 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`

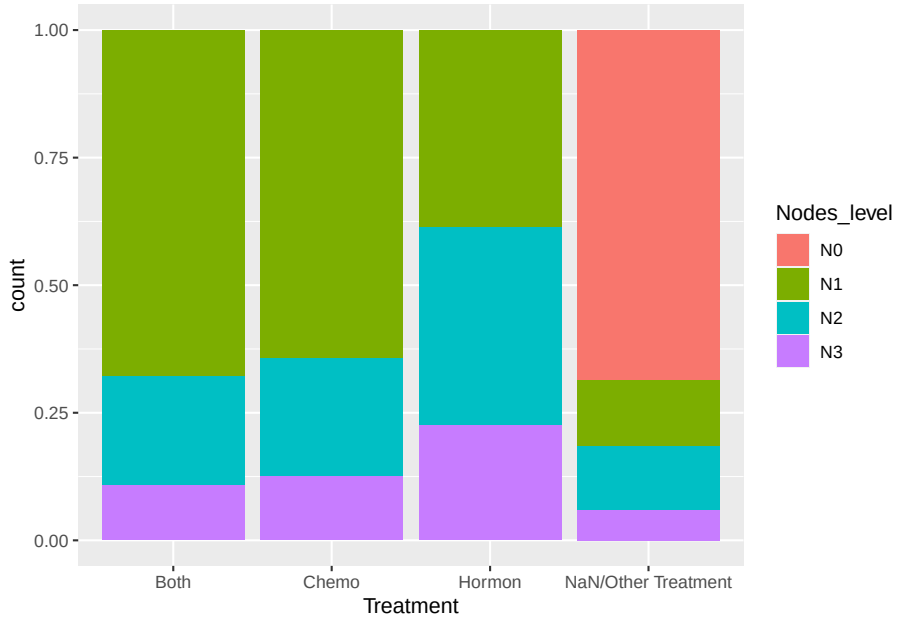
```
ggplot(data = rotterdam, aes(x = Treatment, fill = size)) +
  geom_bar(position = "fill")
```



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 obvious distinction in the treatment taken among different sizes of tumor as we can see from the plot above.

`Treatment` vs. `Nodes_level`

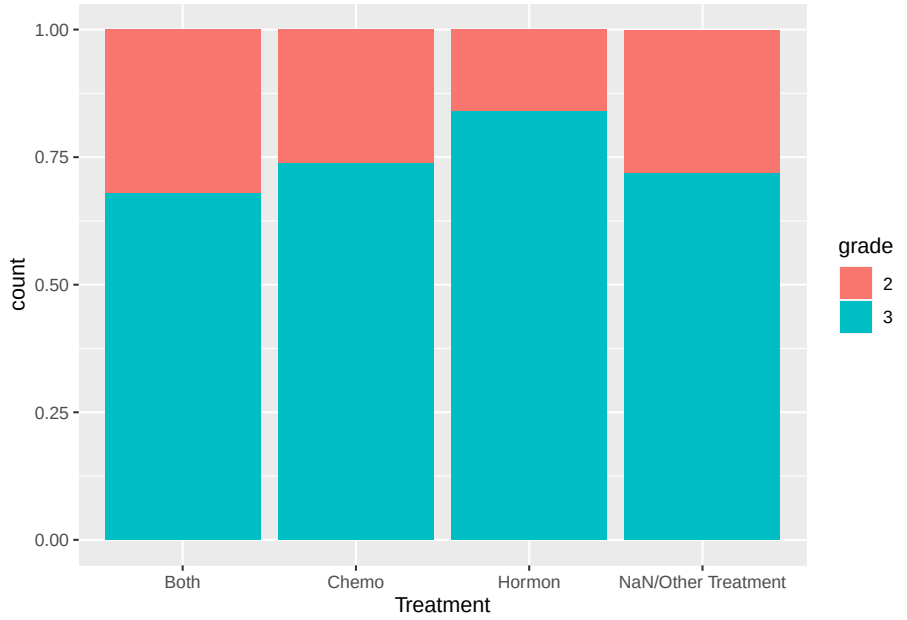
```
ggplot(data = rotterdam, aes(x = Treatment, fill = Nodes_level)) +
  geom_bar(position = "fill")
```



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.

`Treatment` vs. `grade`

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



From the above we can see that for grade 2 and grade 3 breast cancer, there's not too much difference in the treatment a patient takes.

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 grade III breast cancer, which is pretty severe. This would give a more 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 Analysis

3.1 Notation

In survival analysis, there is an important concept called "censoring". Censoring is a kind of missing data problem where an event is not completely observed, for reasons like termination of study or loss of communication with the participant. And censoring will cause the observed time data inaccurate. For instance, for left-censored data, we don't know the exact value of the data point, but only an upper bound for it; And for right-censored data, we only knows a lower bound for it.

In our dataset `rotterdam`, for the time variable `dtime` and `rtime` that we are curious about, `death` and `recur` are the corresponding variable that records if censoring occured. Since if someone is dead in this study, this means that the survival time we observed is exact; But for someone is still alive in the end of the study, the survival time we observed is just a lower bound. And similarly, if someone suffers from recurrent breast cancer during the study, our `rtime` should be accurate; But for someone did not encounter recurrence of breast cancer in the study, we are not sure if she will encounter in the future, so the data point is also a lower point.

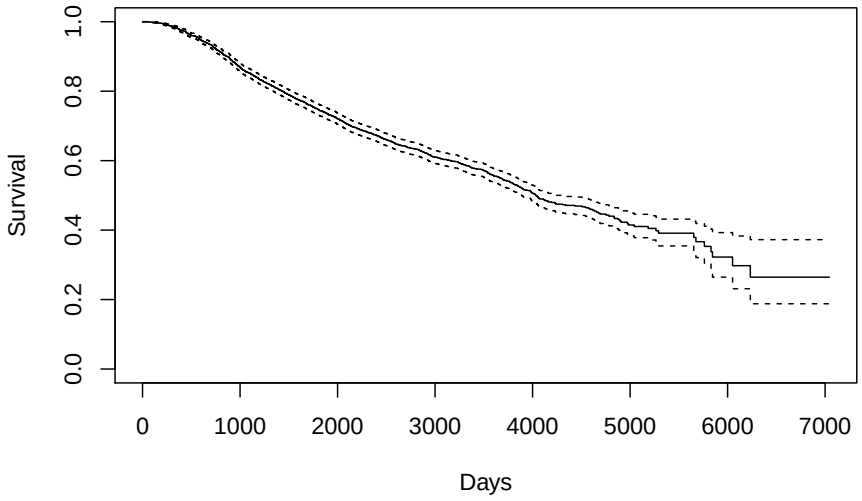
For all the survival analysis below, we will include `death` and `recur` to examine the relationship between `dtime` and `rtime` vs. the diagnostic information and treatment received.

3.2 Kaplan-Miere estimator of the entire dataset

Generally speaking, Kaplan-Miere curve is a non-parametric estimator of the survival function which takes censoring into account. Its y-axis measures $P(X \leq k)$ for various values of k.

3.2.0.1 Death Time

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



Kaplan-Meier curve also helps us finding the median and mean estimate. For median, we can directly find it from the survfit object `KM_d`. For mean, we need to calculate the area under the Kaplan-Meier curve.

```
# median
KM_d
```

```
## Call: survfit(formula = Surv(dtime, death) ~ 1, data = rotterdam)
##
##       n  events  median 0.95LCL 0.95UCL
##    2982    1272    4033    3888    4309
```

```
# mean

# AUCKM stands for "Area Under Curve Kaplan Meier":
AUCKM = function(survobj,duration)
{
base=c(0,summary(survobj)$time,max(duration))
heights=c(1,summary(survobj)$surv)
new=c()
for(i in 1:length(heights)) { new=c(new,(base[i+1]-base[i])*heights[i]) }
c(sum(new))
}

AUCKM(KM_d,rotterdam$dtime)
```

```
## [1] 4099.795
```

The overall mean survival time till death for breast cancer is 4099.795 days, and the overall median survival time till death for breast cancer is 4033 days. Both of them are approximately 7 years.

We can find that the mean and median we calculated above is much longer than the mean and median of the variable `dtime` itself, because a lot of data point are right-censored:

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

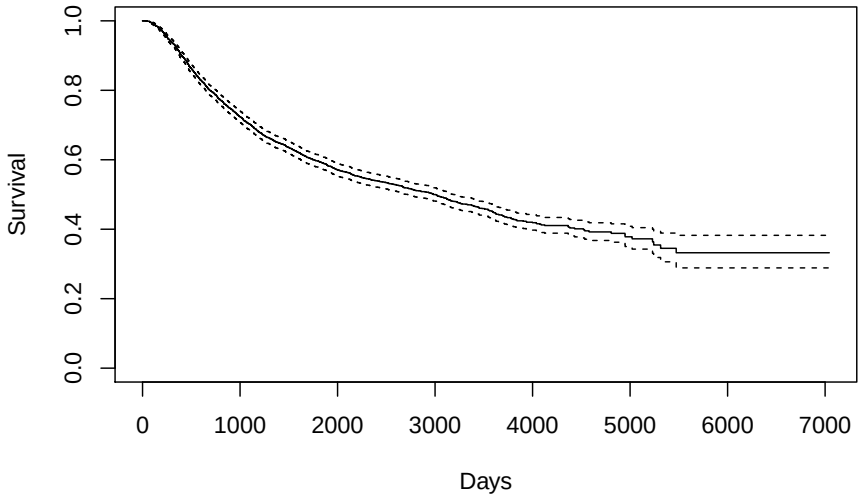
```
## [1] 2605.34
```

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

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

3.2.0.2 Recurrence Time

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



We can find the mean and median in a similar way:

```
# median
KM_r
```

```
## Call: survfit(formula = Surv(rtime, recur) ~ 1, data = rotterdam)
##
##       n  events  median 0.95LCL 0.95UCL
##    2982    1518    2983    2719    3193
```

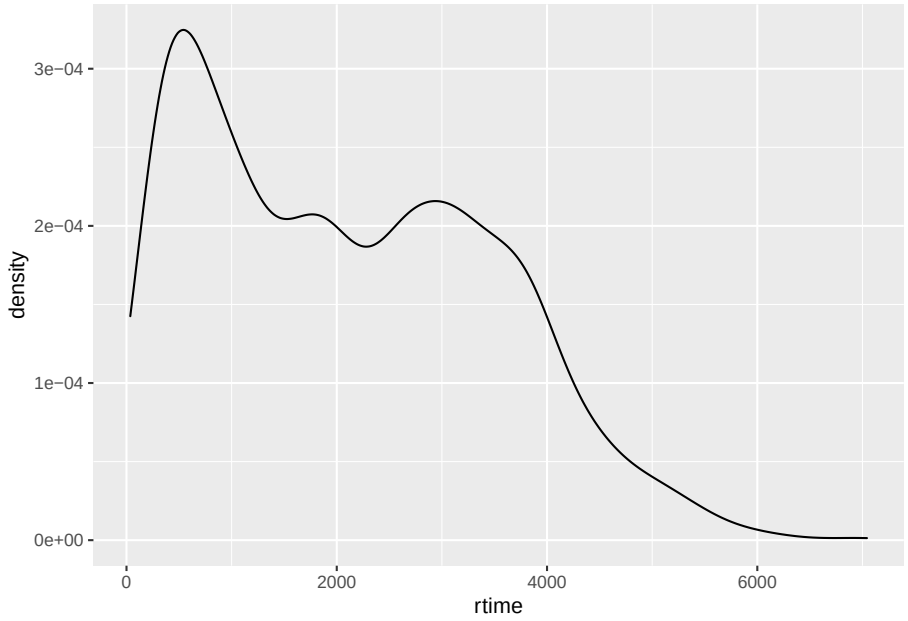
```
# mean

AUCKM(KM_r,rotterdam$rtime)
```

```
## [1] 3588.535
```

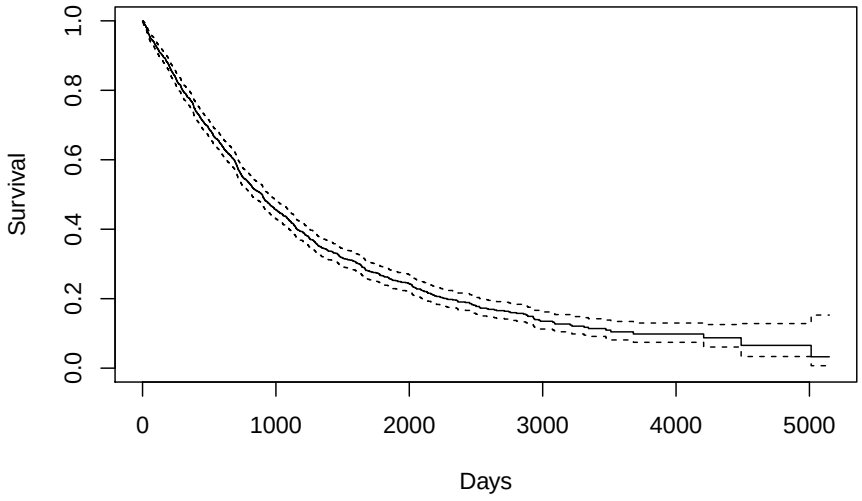
The overall mean survival time till recurrence for breast cancer is 3588.535 days, which is approximately 10 years. The overall median survival time till recurrence for breast cancer is 2983 days, which is approximately 8 years. Notice here our mean is much greater than the median, and this indicates that our `rtime` should be right-skewed.

```
ggplot(rotterdam, aes(x=rtime)) +
   geom_density()
```



3.2.0.3 Survival time after recurrence

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




```
# median
KM_dr
```

```
## Call: survfit(formula = Surv(drecurtime, death) ~ 1, data = rotterdam_recur)
##
##       n  events  median 0.95LCL 0.95UCL
##    1518    1077     894     815     949
```

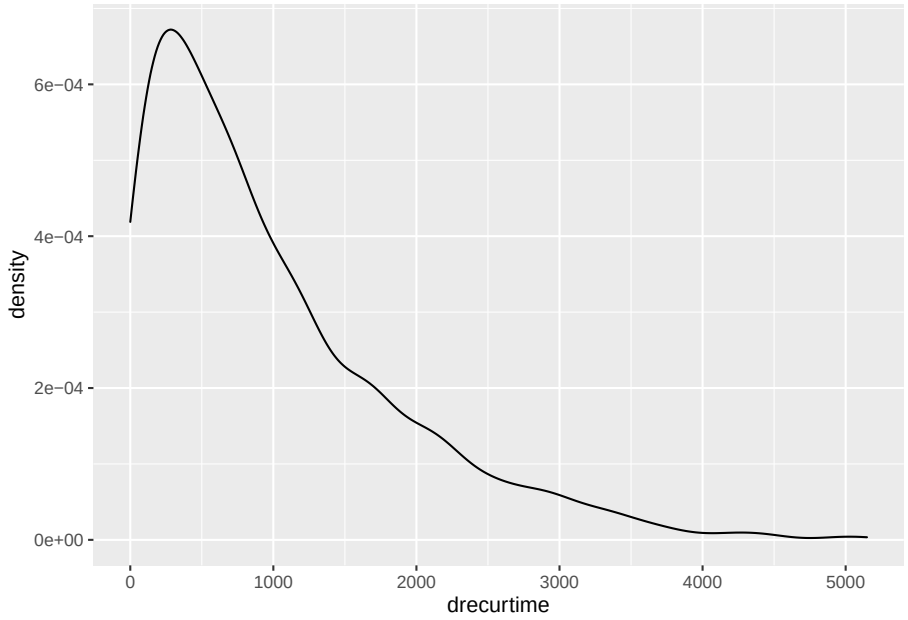
```
# mean

AUCKM(KM_dr,rotterdam_recur$drecurtime)
```

```
## [1] 1407.135
```

The overall mean survival time after rucurrence till death for breast cancer is 1407.135 days, which is approximately a little less than 4 years. The overall median survival time after rucurrence till death for breast cancer is 894 days, which is approximately 2 years and a half. Here, the mean is also greater than the median, also indicating right-skewness:

```
ggplot(rotterdam_recur, aes(x=drecurtime)) +
   geom_density()
```



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

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

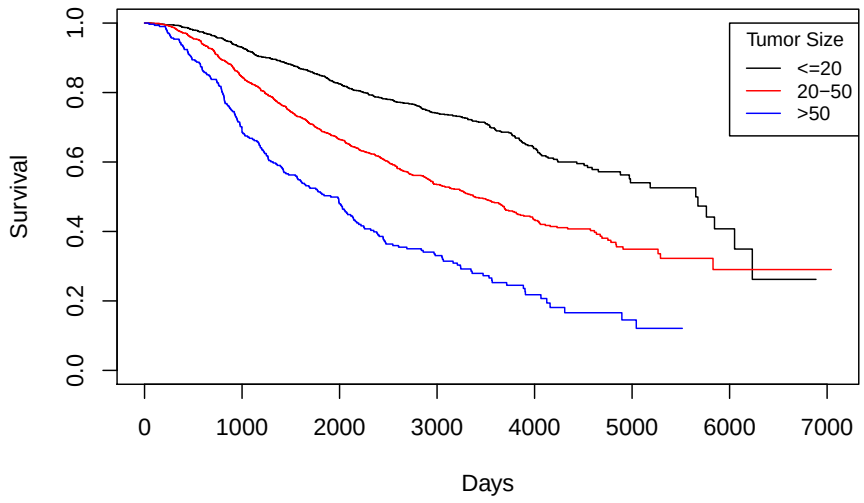
There are 16 variables in the `rotterdam` dataset. Of course we could have fit each variable with a KM estimator, but it would be meaningless to do them all. We will stick to the Diagnostics and Treatment we mentioned in Chapter 2 and fit `size`, `Nodes_level`(we are not using nodes because a Kaplan-Miere estimator does not work well with quantitative variables), `grade` 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`

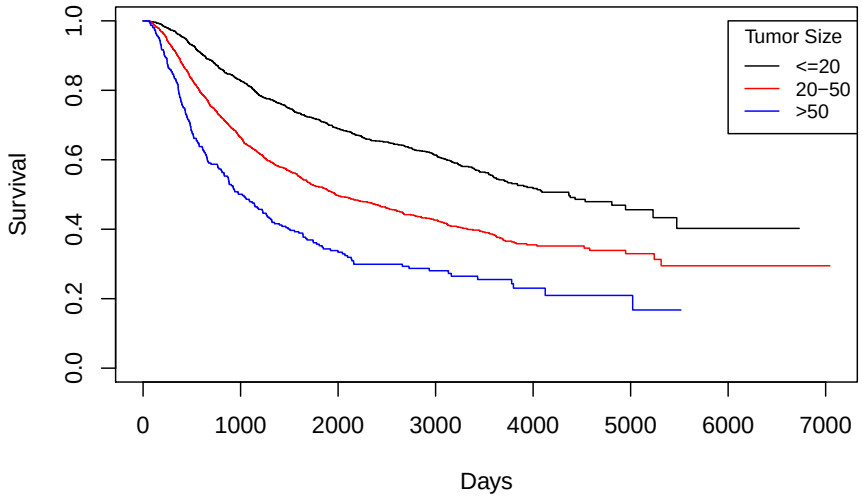
```
KM_None_Death <- survfit(Surv(dtime, death) ~ size, data = rotterdam)
plot(KM_None_Death, conf.type = "plain", col = c("black","red","blue"), xlab="Days", ylab="Surviv
legend(6000, 1, legend=c("<=20", "20-50", ">50"),
```

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



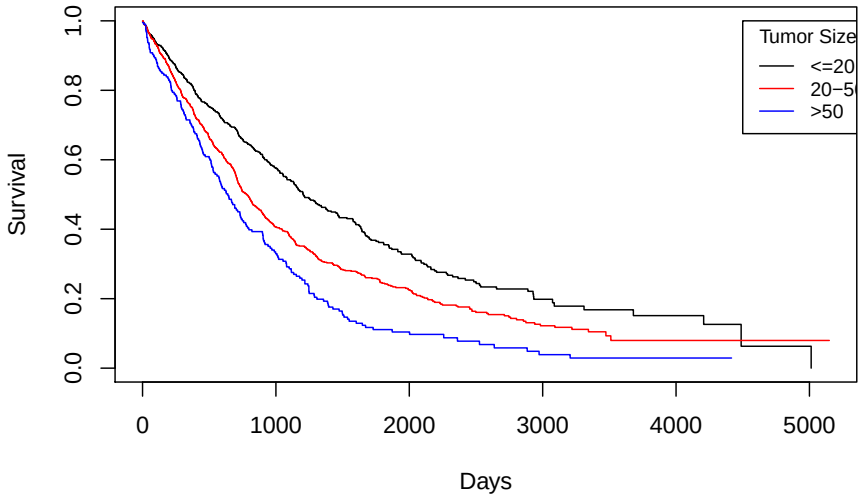
`size` vs. `rtime`

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

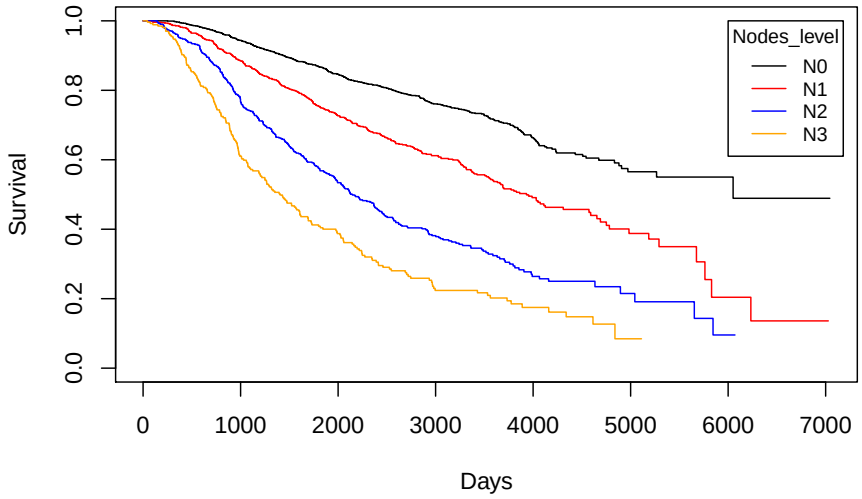


The trend is still the same as patients with smaller tumor size enjoy longer survival of death after recurrence, but the survival time now decreases much faster for all groups.

3.3.2 `Nodes_level` vs. Survival Times

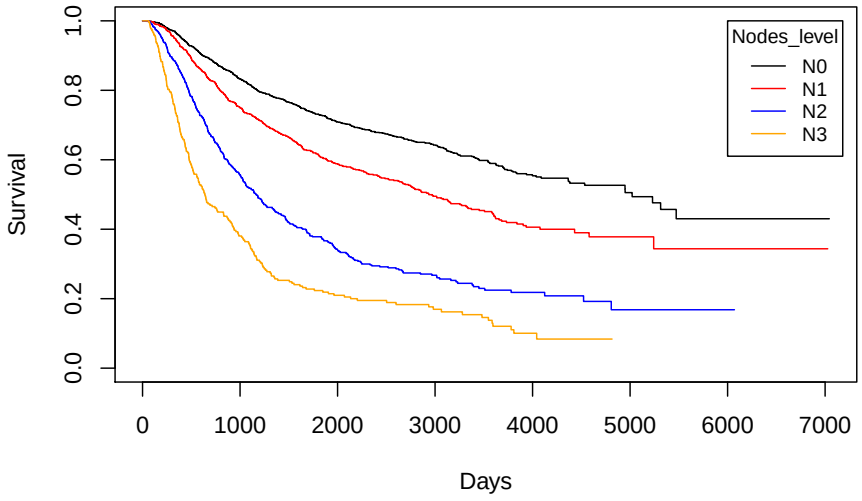
`Nodes_level` vs. `dtime`

```
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

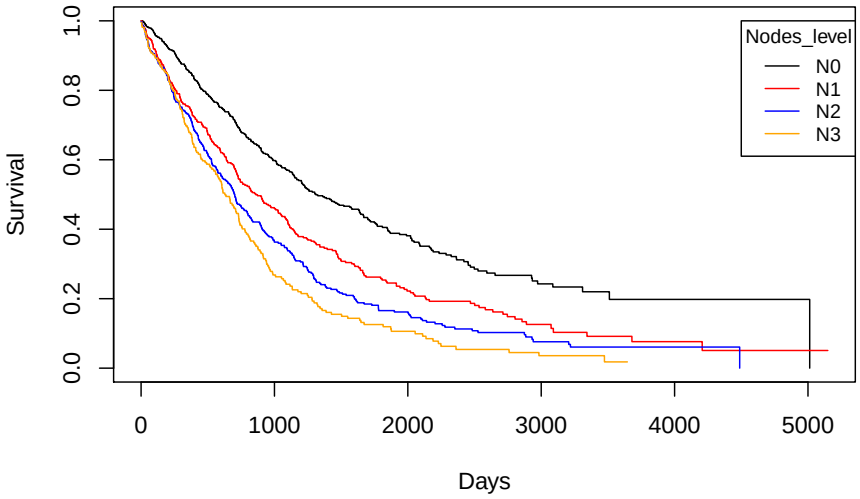
```
KM_None_Recur <- survfit(Surv(rtime, recur) ~ Nodes_level, data = rotterdam)
plot(KM_None_Recur, conf.type = "plain", col = c("black","red","blue","orange"), xlab="Days", yla
legend(6000, 1, legend=c("N0", "N1","N2", "N3"),
       col=c("black", "red", "blue", "orange"), lty=1, cex=0.8,
       title="Nodes_level",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

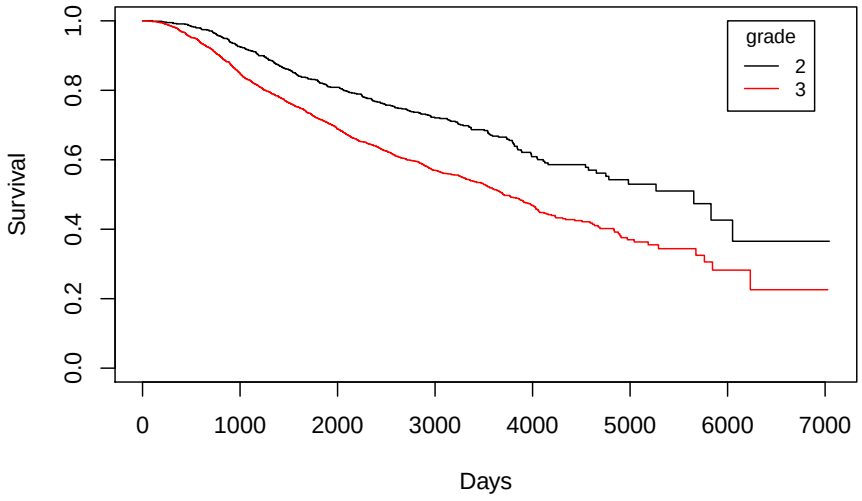


Similarly, the trend is still the same as patients with fewer nodes tested positive enjoy longer survival of death after recurrence, but the survival time now decreases much faster for all groups, and the difference is small in groups `N1`, `N2` and `N3`.

3.3.3 `grade` vs. Survival Times

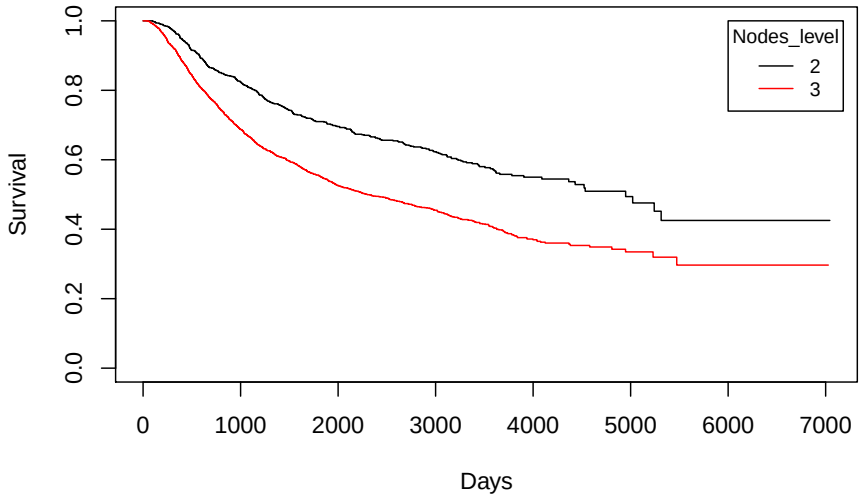
`grade` vs. `dtime`

```
KM_None_Death <- survfit(Surv(dtime, death) ~ grade, data = rotterdam)
plot(KM_None_Death, conf.type = "plain", col = c("black","red"), xlab="Days", ylab="Survival")
legend(6000, 1, legend=c("2", "3"),
       col=c("black", "red"), lty=1, cex=0.8,
       title="grade", text.font=6)
```



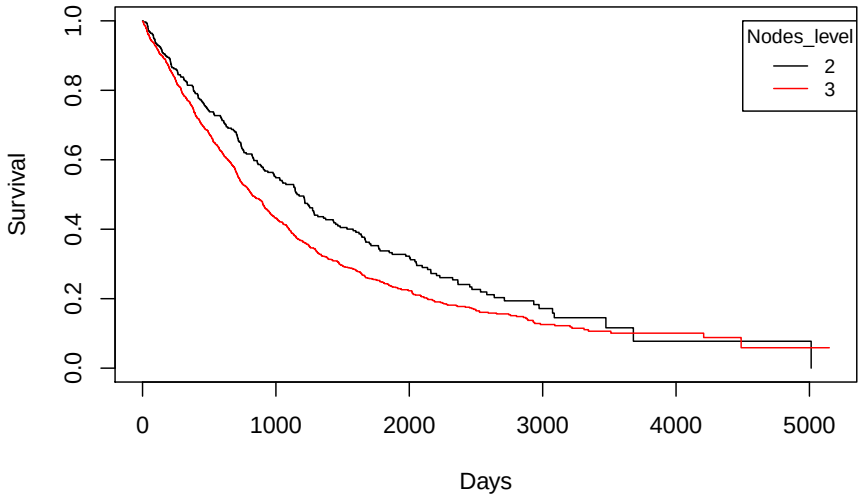
`grade` vs. `rtime`

```
KM_None_Recur <- survfit(Surv(rtime, recur) ~ grade, data = rotterdam)
plot(KM_None_Recur, conf.type = "plain", col = c("black","red"), xlab="Days", ylab="Su
legend(6000, 1, legend=c("2", "3"),
       col=c("black", "red"), lty=1, cex=0.8,
       title="Nodes_level", text.font=6)
```



grade vs. drecurtime

```
KM_None_drecur <- survfit(Surv(drecurtime, death) ~ grade, data = rotterdam_recur)
plot(KM_None_drecur, conf.type = "plain", col = c("black","red"), xlab="Days", ylab="Survival")
legend(4500, 1, legend=c("2", "3"),
       col=c("black", "red"), lty=1, cex=0.8,
       title="Nodes_level", text.font=6)
```

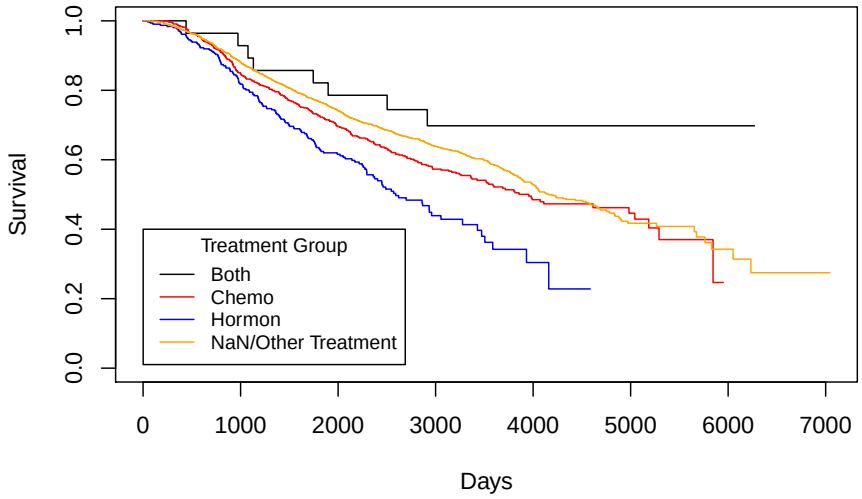


Similarly, we find that patients with grade 2 breast cancer show clear distinction in both the survival of death and recurrence. However, such distinct becomes less obvious for the survival of death after recurrence has occurred.

3.3.4 `Treatment` vs. Survival Times

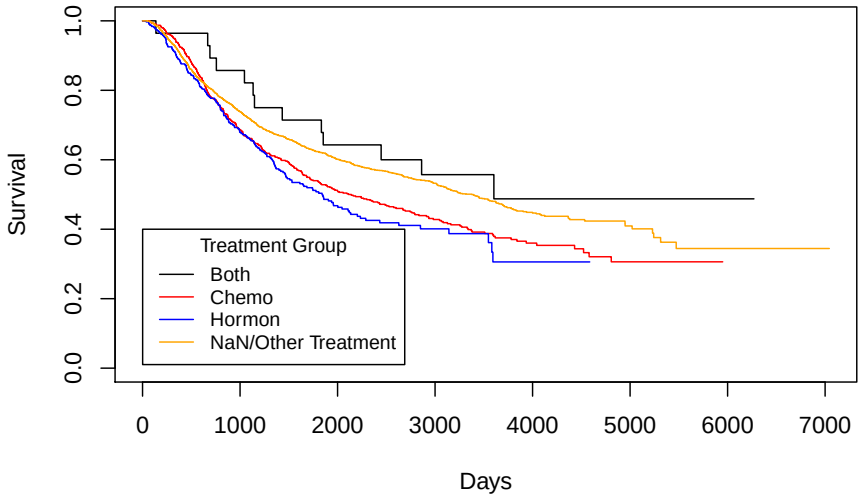
`Treatment` vs. `dtime`

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



Treatment vs. rtime

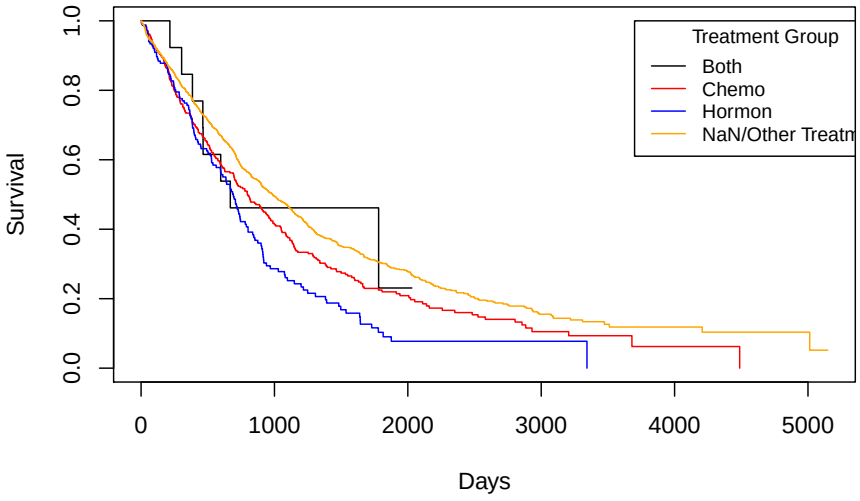
```
KM_Treatment_Recur <- survfit(Surv(rtime, recur) ~ Treatment, data = rotterdam)
plot(KM_Treatment_Recur, conf.int = FALSE, col = c("black", "red", "blue", "orange"), xlab="Days"
legend(1, 0.4, legend=c("Both", "Chemo","Hormon", "NaN/Other Treatment"),
       col=c("black", "red", "blue", "orange"),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



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_re
plot(KM_Treatment_drecur, conf.int = FALSE, col = c("black", "red", "blue", "orange"),
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`, `grade`, `age` (we have shown in Chapter 2 that age is a confounder for categories in Treatment).

3.4.1 User-defined Cox-Snell function

If we fit Exponential model on the survival time, we have the function $S(t) = e^{-\lambda t}$, and we can transform the survival function into an expression for the "complementary log-log":

$$
\begin{aligned}
S(t) &= e^{-\lambda t} \\
\log[S(t)] &= -\lambda t \\
-\log[S(t)] &= \lambda t \\
\log\big(-\log[S(t)]\big) &= \log\lambda + \log t
\end{aligned}
$$

Then, if we let $y = \log\big(-\log[S(t)]\big)$, $m = 1$, $b = \log(\lambda)$, and $x = \log(t)$, then we can view the equation in the form $y = mx + b = 1 * x + b$.

This means that if Exponential model is adequate, its complementary log-log plot should be a linear line with slope=1 and intercept=0.

And for any other parametric model, we can define a Cox-Snell residual as $CS_i = -log(\hat{S}_p(t_i|x_i))$ for $i = 1, \ldots, n$, and CS_is should behave like data drawn from an Exponential distribution ($\lambda = 1$).

Thus, we can check the adequacy of any model by checking the Exponentialness of its Cox-Snell residuals. We can graph CS_i vs $log(CS_i)$, and the graph should also look linear with slope =1 and intercept=0, if the parametric model is adequate for the data.

```
# The Cox-Snell function takes as inputs
# 1. A vector of Cox-Snell residuals created by the user based on the model being eval
# 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')
}
```

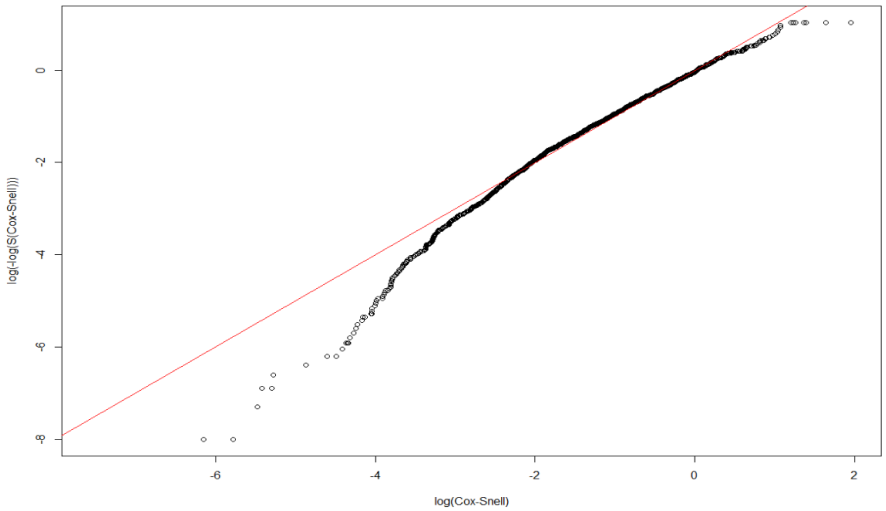
3.4.2 Exponential and Weibull

We first tried to examine the adquacy of Exponential and Weibull model, but none of them are adquate for our dataset:

1. `dtime, death ~ Treatment + size + nodes + grade + age` under Exponential:

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

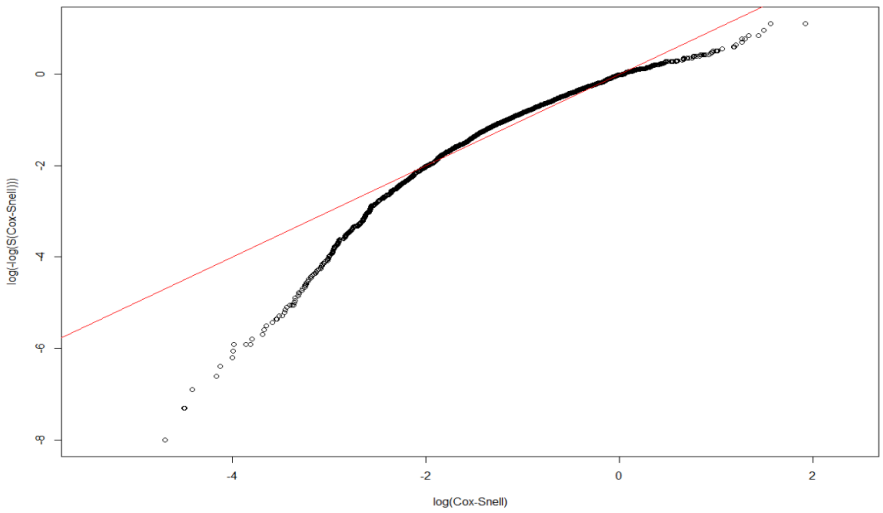
```
## Call:
## survreg(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##     grade + age, data = rotterdam, dist = "exponential")
##
## Coefficients:
##                   (Intercept)                TreatmentChemo
##                   10.72043480                   -0.51333347
##              TreatmentHormon TreatmentNaN/Other Treatment
##                   -0.41382108                   -0.45674929
##                     size20-50                       size>50
##                   -0.44229181                   -0.76874022
##                         nodes                        grade3
##                   -0.06922296                   -0.32925464
##                           age
##                   -0.01424763
##
## Scale fixed at 1
##
## Loglik(model)= -12125.2   Loglik(intercept only)= -12360.4
##  Chisq= 470.37 on 8 degrees of freedom, p= <2e-16
## n= 2982
```



2. `rtime, recur ~ Treatment + size + nodes + grade + age` under Exponential:

```
Rexp <- survreg(Surv(rtime, recur) ~ Treatment + size + nodes + grade + age, dist='expo
Rexp
```

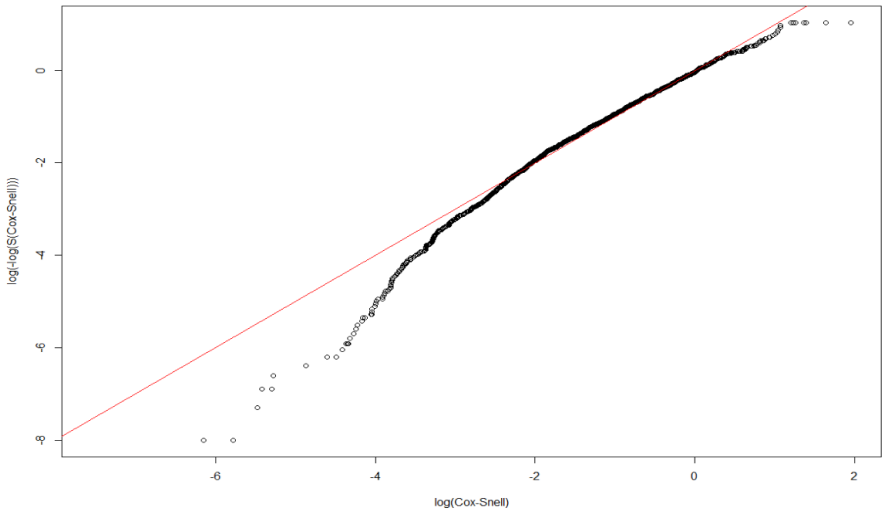
```
## Call:
## survreg(formula = Surv(rtime, recur) ~ Treatment + size + nodes +
##     grade + age, data = rotterdam, dist = "exponential")
##
## Coefficients:
##                   (Intercept)                TreatmentChemo
##                  9.081159515                  -0.373811470
##              TreatmentHormon TreatmentNaN/Other Treatment
##                 -0.507332960                  -0.485387357
##                    size20-50                       size>50
##                 -0.372059978                  -0.654240460
##                         nodes                        grade3
##                 -0.080821343                  -0.393920415
##                           age
##                  0.008095379
##
## Scale fixed at 1
##
## Loglik(model)= -13897.4   Loglik(intercept only)= -14153.7
##  Chisq= 512.6 on 8 degrees of freedom, p= <2e-16
## n= 2982
```



3. `dtime, death ~ Treatment + size + nodes + grade + age` under Weibull:

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

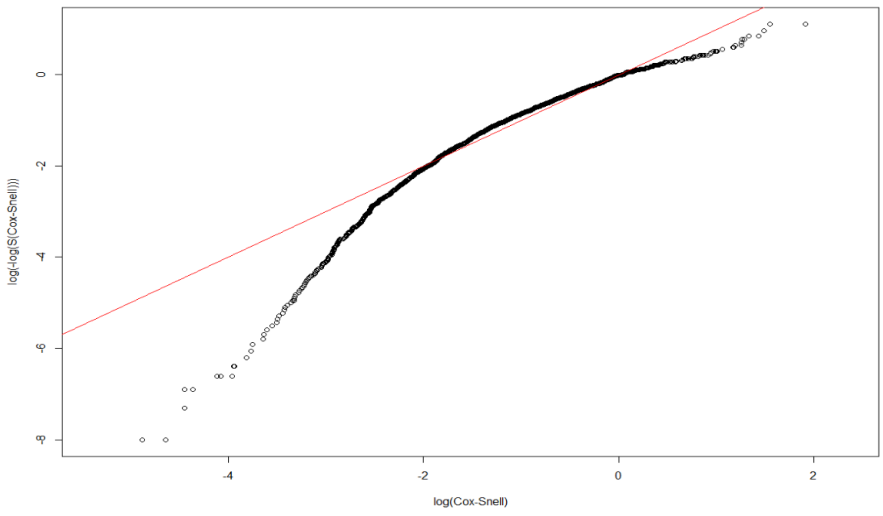
```
## Call:
## survreg(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##     grade + age, data = rotterdam, dist = "weibull")
##
## Coefficients:
##                   (Intercept)                TreatmentChemo
##                   10.10735484                   -0.39504003
##               TreatmentHormon TreatmentNaN/Other Treatment
##                   -0.37941684                   -0.35423500
##                     size20-50                       size>50
##                   -0.33487166                   -0.61672126
##                         nodes                        grade3
##                   -0.05500626                   -0.26093659
##                           age
##                   -0.01130402
##
## Scale= 0.7370449
##
## Loglik(model)= -12055.3   Loglik(intercept only)= -12322.7
##  Chisq= 534.68 on 8 degrees of freedom, p= <2e-16
## n= 2982
```



4. `rtime, recur ~ Treatment + size + nodes + grade + age` under Weibull:

```
Rweibull <- survreg(Surv(rtime, recur) ~ Treatment + size + nodes + grade + age, dist=
Rweibull
```

```
## Call:
## survreg(formula = Surv(rtime, recur) ~ Treatment + size + nodes +
##     grade + age, data = rotterdam, dist = "weibull")
##
## Coefficients:
##                   (Intercept)                TreatmentChemo
##                  9.092880859                  -0.376425745
##              TreatmentHormon TreatmentNaN/Other Treatment
##                 -0.508855839                  -0.489527042
##                    size20-50                       size>50
##                 -0.376473401                  -0.659532060
##                         nodes                        grade3
##                 -0.081625087                  -0.398251643
##                           age
##                  0.008216064
##
## Scale= 1.014145
##
## Loglik(model)= -13897.2   Loglik(intercept only)= -14145.7
##  Chisq= 497.1 on 8 degrees of freedom, p= <2e-16
## n= 2982
```



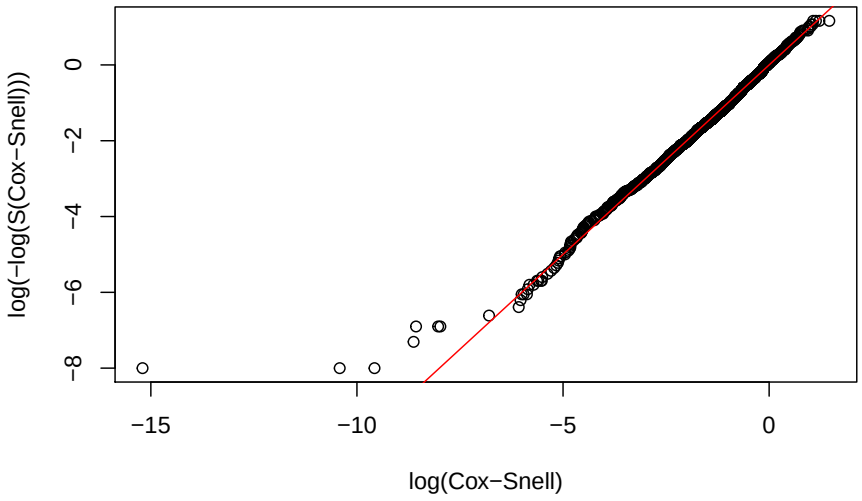
3.4.3 Log-normal models

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

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

```
##
## Call:
## survreg(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##     grade + age, data = rotterdam, dist = "lognormal")
##                                  Value Std. Error      z       p
## (Intercept)                    9.91513    0.27975  35.44 < 2e-16
## TreatmentChemo                -0.42503    0.26292  -1.62   0.106
## TreatmentHormon               -0.31448    0.26896  -1.17   0.242
## TreatmentNaN/Other Treatment -0.41617    0.26011  -1.60   0.110
## size20-50                     -0.34787    0.05108  -6.81 9.8e-12
## size>50                       -0.60916    0.08067  -7.55 4.3e-14
## nodes                         -0.07773    0.00544 -14.29 < 2e-16
## grade3                        -0.30866    0.05490  -5.62 1.9e-08
## age                           -0.01005    0.00197  -5.10 3.3e-07
## Log(scale)                     0.06670    0.02162   3.08   0.002
##
## Scale= 1.07
##
## Log Normal distribution
## Loglik(model)= -12018.2   Loglik(intercept only)= -12286.5
##  Chisq= 536.66 on 8 degrees of freedom, p= 9.5e-111
## Number of Newton-Raphson Iterations: 4
## n= 2982
```

```
CS_LnormD <- -log(1 - plnorm(rotterdam$dtime, 9.91512622-0.42502769*(rotterdam$Treatment=="Chemo"
                                                -0.31448004*(rotterdam$Treatment=="Hormon")
                                                -0.41616843*(rotterdam$Treatment=="NaN/Other Tr
                                                -0.34787441*(rotterdam$size=="20-50")
                                                -0.60916461*(rotterdam$size==">50")
                                                -0.07772719*rotterdam$nodes
                                                -0.30866307*(rotterdam$grade=="3")
                                                -0.01004650*rotterdam$age,
                                        1.06897))
# Make appropriate graph using CoxSnell function
CoxSnell(CS_LnormD, rotterdam$death)
```



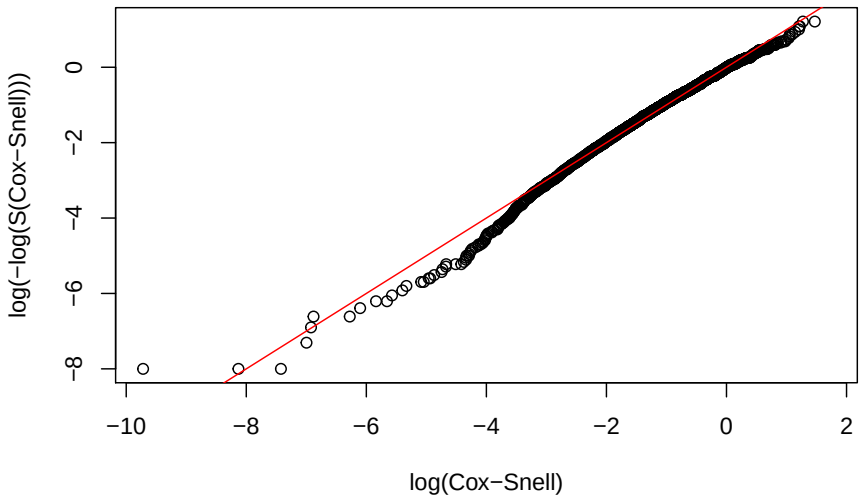

```
Rlnorm <- survreg(Surv(rtime, recur) ~ Treatment + size + nodes + grade + age, dist='l
summary(Rlnorm)
```

```
##
## Call:
## survreg(formula = Surv(rtime, recur) ~ Treatment + size + nodes +
##     grade + age, data = rotterdam, dist = "lognormal")
##                                  Value Std. Error      z       p
## (Intercept)                    8.78912    0.30877  28.46 < 2e-16
## TreatmentChemo                -0.36139    0.29098  -1.24 0.21425
## TreatmentHormon               -0.41809    0.29943  -1.40 0.16263
## TreatmentNaN/Other Treatment -0.58078    0.28741  -2.02 0.04331
## size20-50                     -0.41866    0.05992  -6.99 2.8e-12
## size>50                       -0.67179    0.09845  -6.82 8.9e-12
## nodes                         -0.10590    0.00662 -16.00 < 2e-16
## grade3                        -0.43987    0.06470  -6.80 1.1e-11
## age                            0.00892    0.00238   3.75 0.00018
## Log(scale)                     0.28360    0.01980  14.33 < 2e-16
##
## Scale= 1.33
##
## Log Normal distribution
## Loglik(model)= -13780.4   Loglik(intercept only)= -14045.8
##  Chisq= 530.65 on 8 degrees of freedom, p= 1.9e-109
```

```
## Number of Newton-Raphson Iterations: 4
## n= 2982
```

```
CS_LnormR <- -log(1 - plnorm(rotterdam$rtime, 8.78912009-0.36138607*(rotterdam$Treatment=="Chemo"
                                                -0.41809135*(rotterdam$Treatment=="Hormon")
                                                -0.58077716*(rotterdam$Treatment=="NaN/Other Tr
                                                -0.41865655*(rotterdam$size=="20-50")
                                                -0.67178824*(rotterdam$size==">50")
                                                -0.10590430*rotterdam$nodes
                                                -0.43987388*(rotterdam$grade=="3")
                                                +0.00892277*rotterdam$age,
                                        1.327904))

# 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` + `grade` + `age`.

3.4.3.1 Interpretation of the coefficients

Lognormal is an example of a class of models called Accelerated Failure Time (AFT) models, in which every 1-unit increase in x_i is associated with a scaling of

time by e^{c_i}, where c is the coefficient of x in the regression model. The quantity e^{c_i} is referred to as a time ratio (TR), and thus each c_i represents a log(TR).

Thus, the results of our lognormal model can be interpreted in the following way:

```
exp(-0.42502769) # chemo
```

3.4.3.1.1 For `dtime`:

```
## [1] 0.6537517
```

```
exp(-0.3144800) # hormon
```

```
## [1] 0.7301685
```

```
exp(-0.41616843) # NaN/Other Treatment
```

```
## [1] 0.6595692
```

```
exp(-0.34787441) # size 20-50
```

```
## [1] 0.7061876
```

```
exp(-0.60916461) # size >50
```

```
## [1] 0.543805
```

```
exp(-0.07772719) # nodes
```

```
## [1] 0.9252168
```

```
exp(-0.30866307) # grade 3
```

```
## [1] 0.7344282
```

```
exp(-0.01004650) # age
```

```
## [1] 0.9900038
```

Having chemotherapy, hormon-therapy and NaN/Other Treatment is associated with a scaling of mean survival time by 0.6537517, 0.7301685 and 0.6595692 respectively, compared to having both chemo and hormon therapies. However, the relationships are all insignificant.

Compared to tumor size <=20mm, having tumor size 20-50mm and >50mm will make the expected survival time scaled by 0.7061876 and 0.543805.

For node and age, each extra positive lymph nodes and every 1-year increase in age is associated with a scaling of mean survival time by 0.9252168 and 0.9900038.

And compared to grade II, grade III of cancer cell makes the expected survival time multiplied by 0.7344282.

```
exp(-0.36138607) # chemo
```

3.4.3.1.2 For `rtime`:

```
## [1] 0.69671
```

```
exp(-0.41809135) # hormon
```

```
## [1] 0.6583021
```

```
exp(-0.58077716) # NaN/Other Treatment
```

```
## [1] 0.5594634
```

```
exp(-0.41865655) # size 20-50
```

```
## [1] 0.6579301
```

```
exp(-0.67178824) # size >50
```

```
## [1] 0.5107943
```

```
exp(-0.10590430) # nodes
```

```
## [1] 0.8995107
```

```
exp(-0.43987388) # grade 3
```

```
## [1] 0.6441177
```

```
exp(0.00892277) # age
```

```
## [1] 1.008963
```

Having chemotherapy, hormon-therapy and NaN/Other Treatment is associated with a scaling of mean time till recurrence by 0.69671, 0.6583021 and 0.5594634 respectively, compared to having both chemo and hormon therapies. And the relations here are also insignificant.

Compared to tumor size <=20mm, having tumor size 20-50mm and >50mm will make the expected survival time scaled by 0.6579301 and 0.5107943.

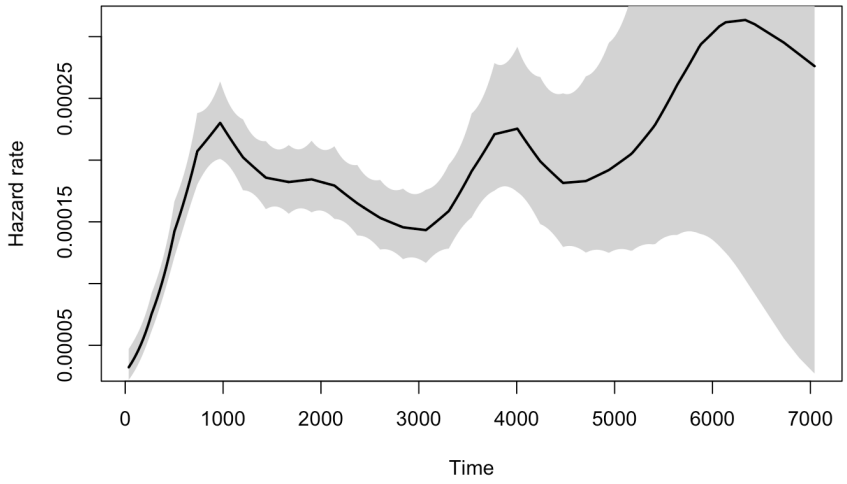
For node and age, each extra positive lymph nodes and every 1-year increase in age is associated with a scaling of mean survival time by 0.8995107 and 1.008963.

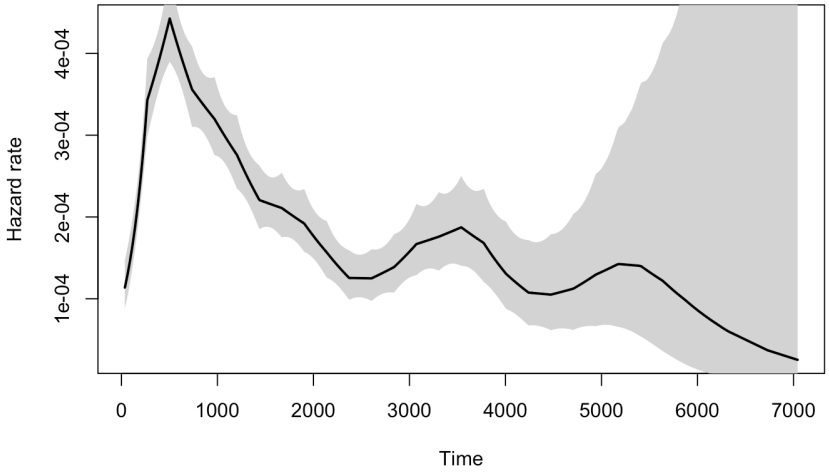
And compared to grade II, grade III of cancer cell makes the expected survival time multiplied by 0.6441177.

3.4.3.2 Why log-normal

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-monotonicity of the hazard function, which is one main characteristic of log-normal model compared to Exponential and Weibull.

The plot below graphs the general hazard functions of our `dtime` and `rtime` in a non-parametric way, and we can find that both of them shows non-monotonicity. Though we cannot graph the exact hazard functions conditioning on the diagnostic and treatment information, the non-parametric hazard can still partially explain the adquacy of log-normal.





3.4.3.3 Positive coefficient of `age`

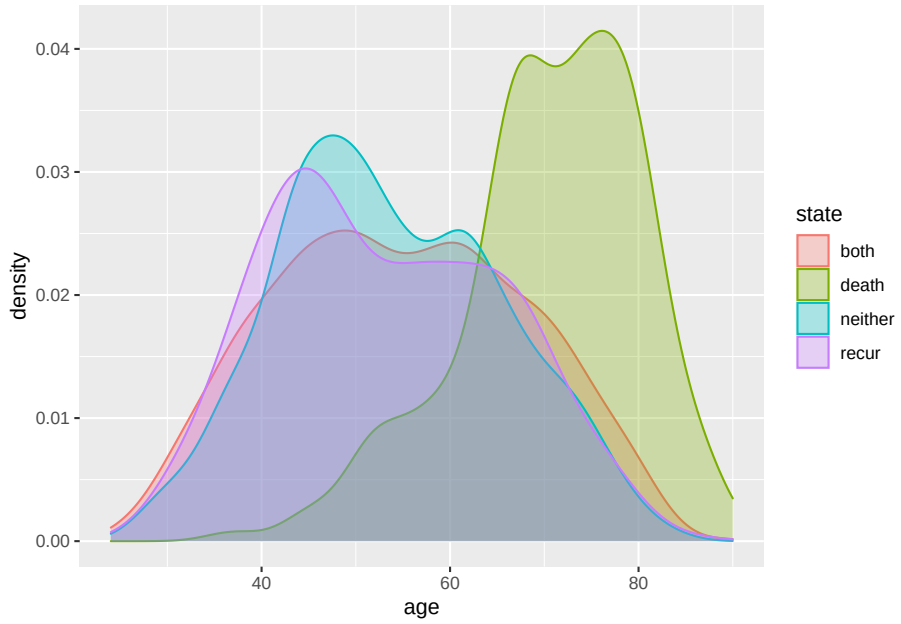
In the `rtime ~ Treatment + size + nodes + age` models, it is surprised to see that the coefficient of `age` is positive, though the value is small. One possible explanation could be the idea of "competing events" and "competing risk".

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 looking for the hazard ratio by building and interpreting the coxph model.

Cox's PH model assumes that the hazard function for any subject can be written as: $h(t) = h_0(t)e^{b_1x_1+b_2x_2+\cdots+b_kx_k}$, where $h_0(t)$ is called a baseline hazard function, and the x_i's are covariates. The Cox PH model assumes that, for any 2 values of a covariate, the hazard ratio (HR) is constant over time.

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` + `grade` + `Treatment`.

3.5.1 Cox-PH model for `dtime`

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

```
## Call:
## coxph(formula = Surv(dtime, death) ~ Treatment + size + nodes +
##     grade + age, data = rotterdam)
##
##                                  coef exp(coef) se(coef)      z        p
## TreatmentChemo               0.544967  1.724551 0.360119  1.513    0.130
## TreatmentHormon              0.488511  1.629887 0.366267  1.334    0.182
## TreatmentNaN/Other Treatment 0.491504  1.634773 0.356997  1.377    0.169
## size20-50                    0.450933  1.569776 0.065260  6.910 4.85e-12
## size>50                      0.818215  2.266451 0.091189  8.973  < 2e-16
## nodes                        0.073329  1.076085 0.004865 15.072  < 2e-16
## grade3                       0.350976  1.420453 0.070141  5.004 5.62e-07
## age                          0.015067  1.015182 0.002552  5.903 3.56e-09
##
## Likelihood ratio test=513.7  on 8 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  4.32  3 0.22855
## size       4.71  2 0.09487
## nodes      3.87  1 0.04915
## grade      2.61  1 0.10637
## age       14.48  1 0.00014
## GLOBAL    26.91  8 0.00073
```

There are two variables, `age` and `nodes`, that have have p-value smaller than 0.05, which indicates the violation of the PH assumption.

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

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

```
## Call:
## coxph(formula = Surv(dtime, death) ~ Treatment + size + strata(nodes) +
##     grade + strata(age), data = rotterdam)
##
##                                  coef exp(coef) se(coef)     z        p
## TreatmentChemo                0.45621   1.57808  0.47456 0.961    0.336
## TreatmentHormon               0.44920   1.56706  0.49904 0.900    0.368
## TreatmentNaN/Other Treatment 0.76097   2.14035  0.48430 1.571    0.116
## size20-50                     0.43163   1.53977  0.08021 5.381 7.40e-08
## size>50                       0.63412   1.88536  0.14221 4.459 8.23e-06
## grade3                        0.38089   1.46358  0.09141 4.167 3.09e-05
##
## Likelihood ratio test=67.3  on 6 df, p=1.459e-12
## n= 2982, number of events= 1272
```

```
cox.zph(m_death_strataage)
```

```
##           chisq df    p
## Treatment 3.198  3 0.36
## size      1.819  2 0.40
## grade     0.635  1 0.43
## GLOBAL    5.791  6 0.45
```

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

For treatments, we can see that Chemotherapy, Hormontherapy and Other Treatment all have higher risks than patients receiving both therapies, with hazard ratio 1.57808, 1.56706 and 2.14035 respectively. However, note that all three p-values are greater than 0.05, which means that we do not have significant evidence for such relationships.

And for size, it is clear that patients with larger tumor size enjoy higher risks, and as the size grows, the risk is also increasing, with hazard ratio 1.53977 and 1.88536 for 20-50mm vs. <=20mm and >50mm vs. <=20mm. Such relationship is significant as we can see from the p-values since they are all less than 0.05.

For grade, we could see that the risk is going to be multiplied by 1.46358, if it goes from grade II to grade III, and the relation is staitstically significant.

3.5.2 Cox-PH model for `rtime`

```
m_recur = coxph(Surv(rtime, recur) ~ Treatment + strata(size) + strata(nodes) + strata(grade) + a
m_recur
```

```
## Call:
## coxph(formula = Surv(rtime, recur) ~ Treatment + strata(size) +
##     strata(nodes) + strata(grade) + age, data = rotterdam)
##
##                                  coef exp(coef)  se(coef)      z        p
## TreatmentChemo                0.280433  1.323703  0.290152  0.967  0.33379
## TreatmentHormon               0.396060  1.485958  0.300374  1.319  0.18732
## TreatmentNaN/Other Treatment  0.817338  2.264464  0.291876  2.800  0.00511
## age                          -0.015377  0.984741  0.002553 -6.023 1.71e-09
##
## Likelihood ratio test=63.27  on 4 df, p=5.96e-13
## n= 2982, number of events= 1518
```

For `rtime`, we found that all the diagnostic information `size`, `nodes` and `grade` obeys the PH assumption, and we can only investigate `Treatment` and `age`.

We can find that patients taking both chemo and hormon therapies still have lowest hazard of breast cancer recurrence, and having chemo only, hormon only or NaN/Other Treatment will multiply the hazard by 1.323703, 1.485958 and 2.264464 respectively. However, the statistical significance is not that strong.

And for age, it's still surprised to see that its coefficient is negative, indicating that every 1-year increase in age will make the hazard of breast cancer recurrence by 0.984741. The reason behind this should be the same as the idea of competing risk we mentioned above.

3.6 Summary

Based on all the survival analysis above, we found that different treatment ways actually does not differ significantly in extending patients' survival time or time till recurrence, and the main factors that affect patients' life are the diagnostic information: tumor size, number of positive lymph nodes and tumor grade.

This reminds us the importance of taking screening every three years. Current medical level still fails to cure breast cancer, but with early diagnosis, patients' lives can be extended by a lot.

Chapter 4

Random Survival Forest

Another point of interest might be how could we predict a patients survival based on some of his/ her diagnostics and treatment records. In order to do this, we will fit a random survival forest model to make predictions.

4.1 dtime

```
set.seed(453)
# Sample the data and create a training subset.
train <- sample(1:nrow(rotterdam), round(nrow(rotterdam) * 0.80))
# Train the model.
rotterdam.grow <- rfsrc(Surv(dtime, death) ~ Treatment + size + nodes + age + grade, rotterdam[tr
# Test the model.
rotterdam.pred <- predict(rotterdam.grow, rotterdam[-train , ])
# Compare the results.
print(rotterdam.grow)
```

```
##                          Sample size: 2386
##                     Number of deaths: 1022
##                      Number of trees: 100
##            Forest terminal node size: 15
##        Average no. of terminal nodes: 110.28
## No. of variables tried at each split: 3
##               Total no. of variables: 5
##        Resampling used to grow trees: swor
##     Resample size used to grow trees: 1508
##                             Analysis: RSF
```

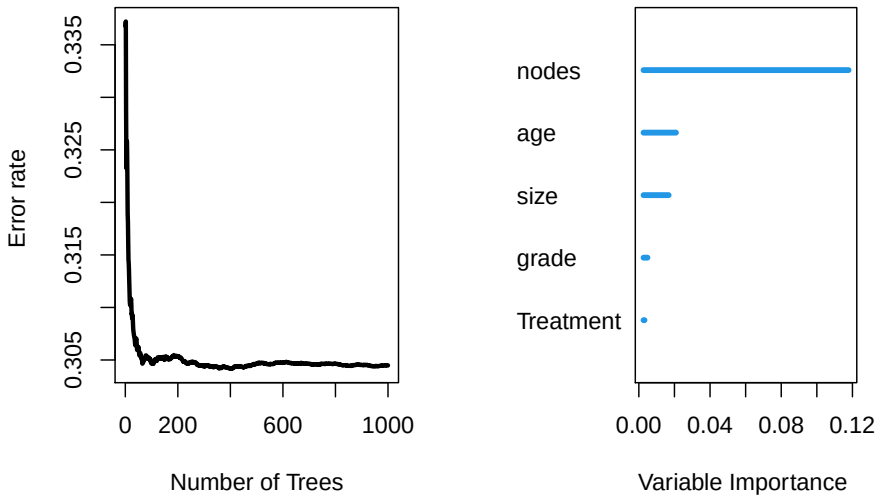
```
##                                 Family: surv
##                         Splitting rule: logrank *random*
##         Number of random split points: 10
##                               Error rate: 31.87%
```

```
print(rotterdam.pred)
```

```
##   Sample size of test (predict) data: 596
##        Number of deaths in test data: 250
##                  Number of grow trees: 100
##   Average no. of grow terminal nodes: 110.28
##           Total no. of grow variables: 5
##        Resampling used to grow trees: swor
##     Resample size used to grow trees: 377
##                                Analysis: RSF
##                                  Family: surv
##                  Test set error rate: 27.43%
```

As we can see from the RSF result, our test set error rate is 27.43%, which is not very ideal.

```
plot.rfsrc(rfsrc(Surv(dtime, death) ~ age + Treatment + size + nodes + grade, rotterdar
```



```
##
##             Importance   Relative Imp
## nodes           0.1178         1.0000
## age             0.0209         0.1771
## size            0.0168         0.1423
## grade           0.0049         0.0412
## Treatment       0.0026         0.0223
```

As we can see, in our fit of RSF, the most important variable is `nodes` and the second most important variable is `age`. Both variables are quantitative variables.

4.2 rtime

```
set.seed(453)
# Sample the data and create a training subset.
train <- sample(1:nrow(rotterdam), round(nrow(rotterdam) * 0.80))
# Train the model.
rotterdam.grow <- rfsrc(Surv(rtime, recur) ~ Treatment + size + nodes + age + grade, rotterdam[tr
# Test the model.
rotterdam.pred <- predict(rotterdam.grow, rotterdam[-train , ])
# Compare the results.
print(rotterdam.grow)
```

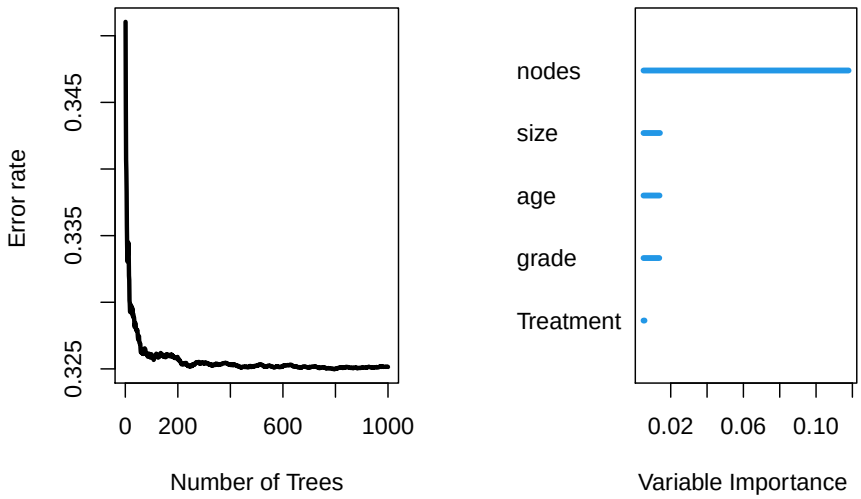
```
##                          Sample size: 2386
##                       Number of deaths: 1211
##                        Number of trees: 100
##              Forest terminal node size: 15
##          Average no. of terminal nodes: 111.27
## No. of variables tried at each split: 3
##                 Total no. of variables: 5
##          Resampling used to grow trees: swor
##       Resample size used to grow trees: 1508
##                                Analysis: RSF
##                                  Family: surv
##                        Splitting rule: logrank *random*
##          Number of random split points: 10
##                              Error rate: 33.28%
```

```
print(rotterdam.pred)
```

```
##   Sample size of test (predict) data: 596
```

```
##         Number of deaths in test data: 307
##                       Number of grow trees: 100
##    Average no. of grow terminal nodes: 111.27
##            Total no. of grow variables: 5
##         Resampling used to grow trees: swor
##      Resample size used to grow trees: 377
##                                Analysis: RSF
##                                  Family: surv
##                     Test set error rate: 30.23%
```

```
plot.rfsrc(rfsrc(Surv(rtime, recur) ~ age + Treatment + size + nodes + grade, rotterdar
```




```
##
##              Importance   Relative Imp
## nodes            0.1179         1.0000
## size             0.0141         0.1192
## age              0.0138         0.1173
## grade            0.0137         0.1161
## Treatment        0.0050         0.0426
```

Similarly, RSF for predicting `rtime` also have error rate around 30%, and the most important factors are `node`, `size` and `age`.

Though the prediction from RSF is not that accurate, but its variable importance ranking gives us very similar information to Chapter 3, that the diagnostic values and age are the most significant part that decides patients' lives.

Chapter 5

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, we have seen that the general trend is: the larger the size of the tumor, the larger the number of lymph nodes tested positive, and the more abnormal looking cancer cell(higher cancer grade), the shorter a patient's survival might be, for both survival of death and recurrence. 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.

5.1 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, number of lymph nodes tested positive and somewhat incomplete metastasis, as we do not have information on patients with grade I cancer. This might lead to an overall pessimistic conclusion about survival times of breast cancer patients. For future studies, we think that it would be important to include grade I breast cancer incidence, as it is crucial to a more accurate estimate of patients' survival time.

- 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.

- In Chapter 4, we digged a little bit into machine learning related topic using survival data. However, the result was not very satisfactory and we could not find a proper way to make predictions parametrically. For future studys, exploring the possibility of predicting survival using censored data might be of interest.

5.2 Reference

- "Breast Cancer - Stages." Cancer.Net, 14 Aug. 2020, https://www.cancer.net/cancer-types/breast-cancer/stages.

- "Breast Cancer Grades" American Cancer Society, 20 Sep. 2019, https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-grades.html.

- "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).