



**Survival Analysis of
Breast Cancer Patients**
PSTAT 175 Final Project

Yuchan Guo, Zhuoqi Niu, Kelly Wang, David Zhang
December 6, 2019

Abstract

Breast cancer is the second most common cancer diagnosed in women in the United States (Source: [Mayo Clinic](#)). It is important to investigate what factors are associated with longer survival time after diagnosis. We try to build a Cox Proportional Hazards model with factors that patients understand easily.

Background and Data Source

The data is collected by NKI (Netherlands Cancer Institute), and we downloaded it from the Data World website. It includes information for 272 breast cancer patients. While it records many attributes (more than 1000), we select only six factors, and the survival time and status, because they are most easily understood by the layman.

These are:

- Survival (survival time from first diagnosis of cancer), coded in years.
- Eventdeath (whether they died of cancer, or not (censored)), coded as 1=death, 0=censored.

Three conditions:

- Age (age at diagnosis), coded in years.
- Grade (the grade of cancer), coded as 1, 2, 3, representing three levels of invasive breast cancer:
 - grade 1 – looks most like normal breast cells and is usually slow-growing.
 - grade 2 – looks less like normal cells and is growing faster.
 - grade 3 – looks different to normal breast cells and is usually fast-growing.

Grade is different to the stage of breast cancer.

- Diameter (diameter of the tumor), coded in millimeters.

And three treatments:

- Chemo (whether the patient received chemotherapy), coded as 0=no, 1=yes.
- Hormonal (whether the patient received hormonal therapy), coded as 0=no, 1=yes.
- Amputation (whether the patient received a forequarter amputation: cutting off the arm and shoulder, last resort only), coded as 0=no, 1=yes.

Research Question

The combination of condition and treatment is according to common sense: a more serious condition, or lack of treatment, may be associated with shorter survival time. We want to see if given these three conditions, how effective the treatments would be. Also, we wish to predict a survival time, or answer the question “how long do I have”, so to speak.

Data Exploration

```

      age      survival
Min.   :26.00   Min.    : 0.7118
1st Qu.:40.75   1st Qu.: 5.4997
Median :45.00   Median : 7.3593
Mean   :44.05   Mean    : 8.0806
3rd Qu.:49.00   3rd Qu.:10.5127
Max.   :53.00   Max.    :18.3409
$chemo
x
  0    1
165 107

$hormonal
x
  0    1
236  36

$amputation
x
  0    1
152 120

$grade
x
  1    2    3
 71   95 106

```

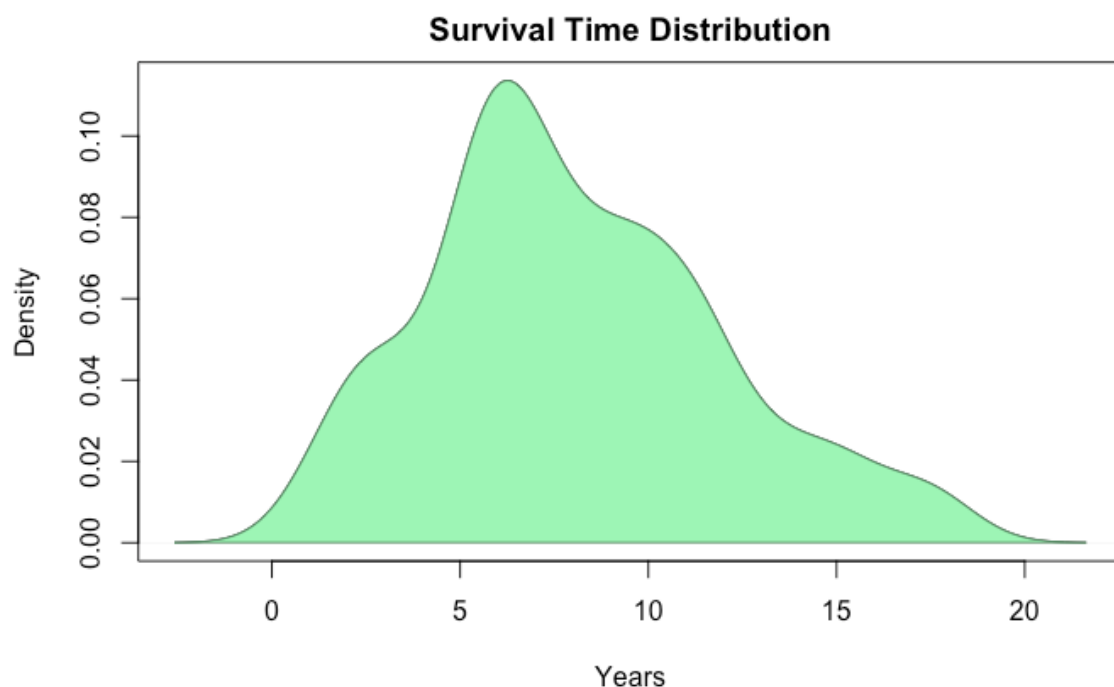
We first want to gain a general understanding of the data. Using the “summary” and “lapply” functions, we find the recorded survival times range from 0.71 years to 18.34 years. The median is 7.36 years, and the average is about 8 years. Now, all of our three treatments are boolean (true/false); and among the three conditions, “age” and “diameter” are numeric, and “grade” is categorical. We often hear that health risks are different for different “age groups”. And “diameter” may also be a condition that can be divided into groups. So, we make “age” and “diam” (as in dataset) categorical.

We group up “age” according to their quantiles, to get “agegroup”. People from age 26.00 to 40.75 (inclusive of 40.75) years are in the first quantile, therefore, in group “1”. People from 40.75 to 49.00 (inclusive of 49.00) years are in the

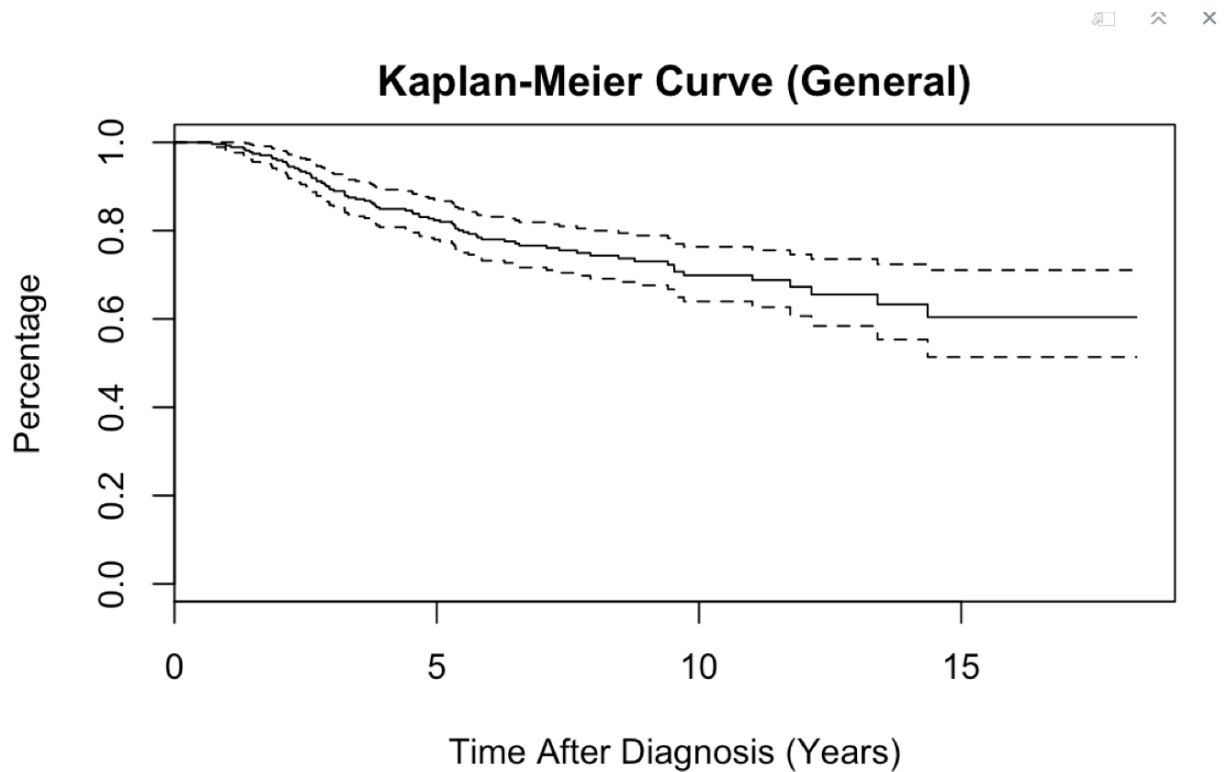
second and third quantiles. These form group “2”. Anyone greater than 49.00 years is in the 4th quantile, therefore in group “3”.

Similarly, we group the “diam” according to their quantiles, to create “diamgroup”. The first quantile is tumor smaller than (inclusive of) 15 millimeters, and we put those observations in “diamgroup” 1. Observations with tumors greater than 15 but less than or equal to 30 millimeters are in the 2nd and 3rd quantiles, which form “diamgroup” 2. The rest of observations with tumor greater than 30 millimeters belong in “diamgroup” 3.

We also plot the distribution of survival time, which illustrates the probability density function (PDF) of survival times. It confirms our previous finding that the 2nd and 3rd quantiles of survival time are located between about 5 and 10.



Then, we do a general Kaplan-Meier curve for all subjects. We see that here, approximately 80% survive at 5 years. The rate given by the United States National Cancer Institute (<https://seer.cancer.gov/statfacts/html/breast.html>) is 89.9%. There may be something particular to NKI subjects that contributes to the difference.

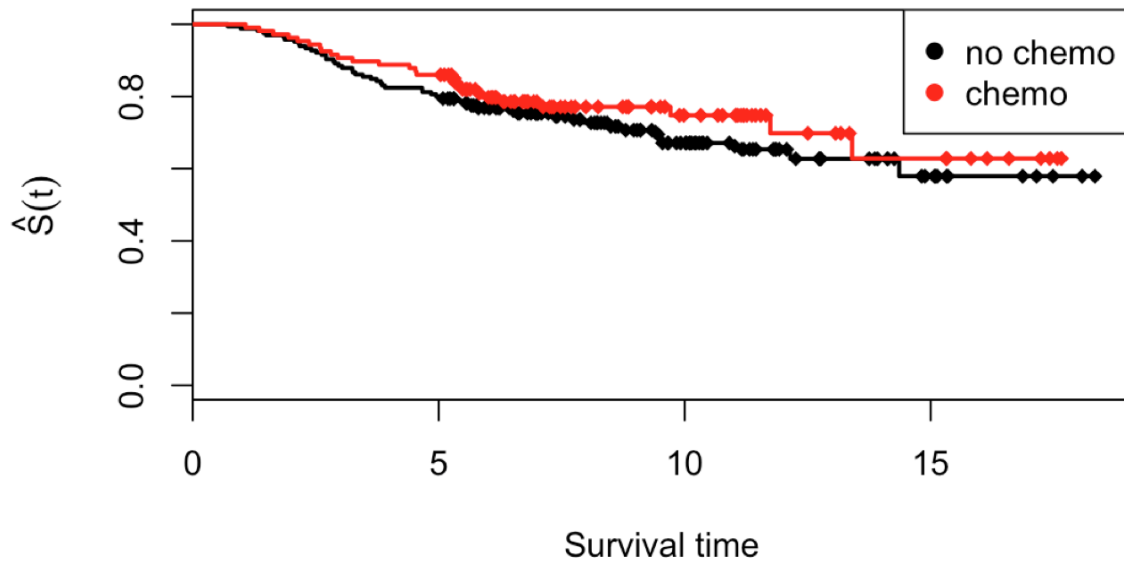


Kaplan-Meier Curves

Next, we plot the Kaplan-Meier survival curves to visually analyze the effects of each covariate: Chemo, hormonal, amputation, grade level, diameter and age group on event death. Looking at the plots below, we can conclude that the three conditions, rather than the three treatments, significantly affect event death.

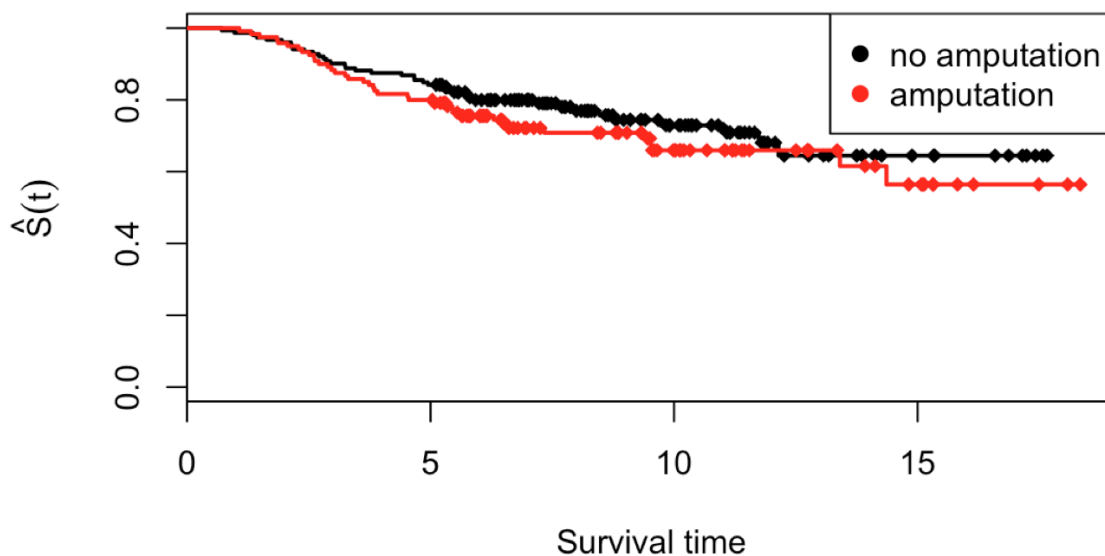
We first plot the three treatments, since the patients and their families would like to know the general effect of treatments, which are, unfortunately, often painful and costly.

KM Curve (Chemo and No Chemo)

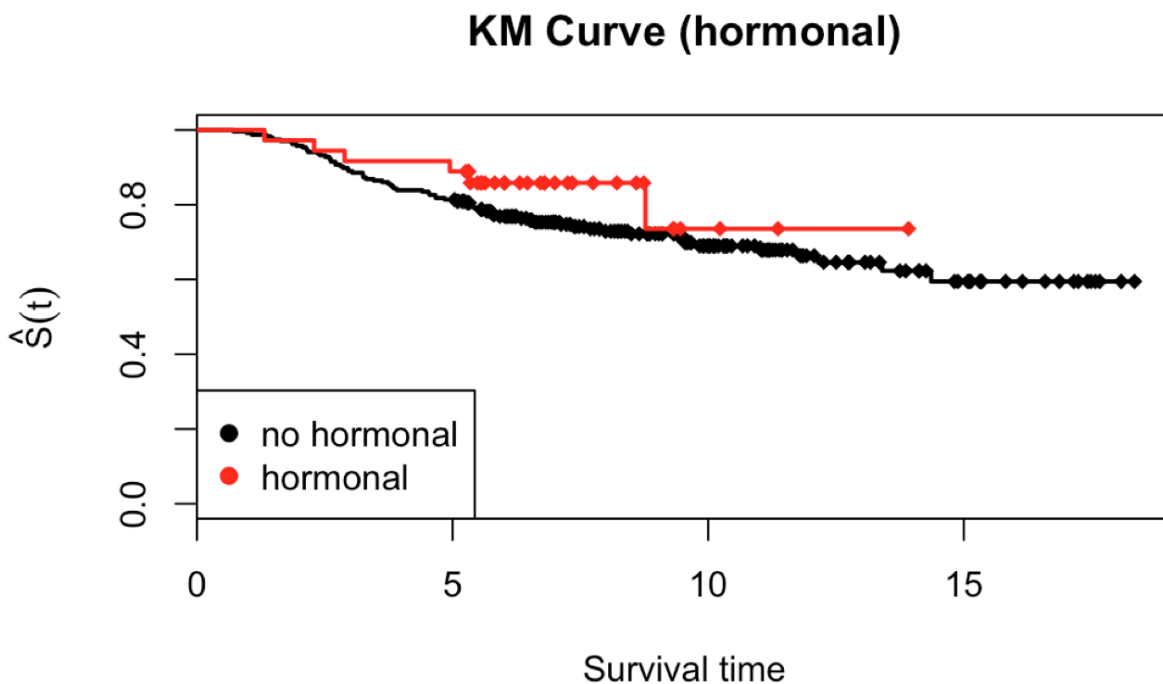


We all know chemotherapy is a commonly used remedy for cancer. However, the curve for “chemo” is only slightly above “no chemo”. There may be no significant difference between the two groups.

KM Curve (Amputation)

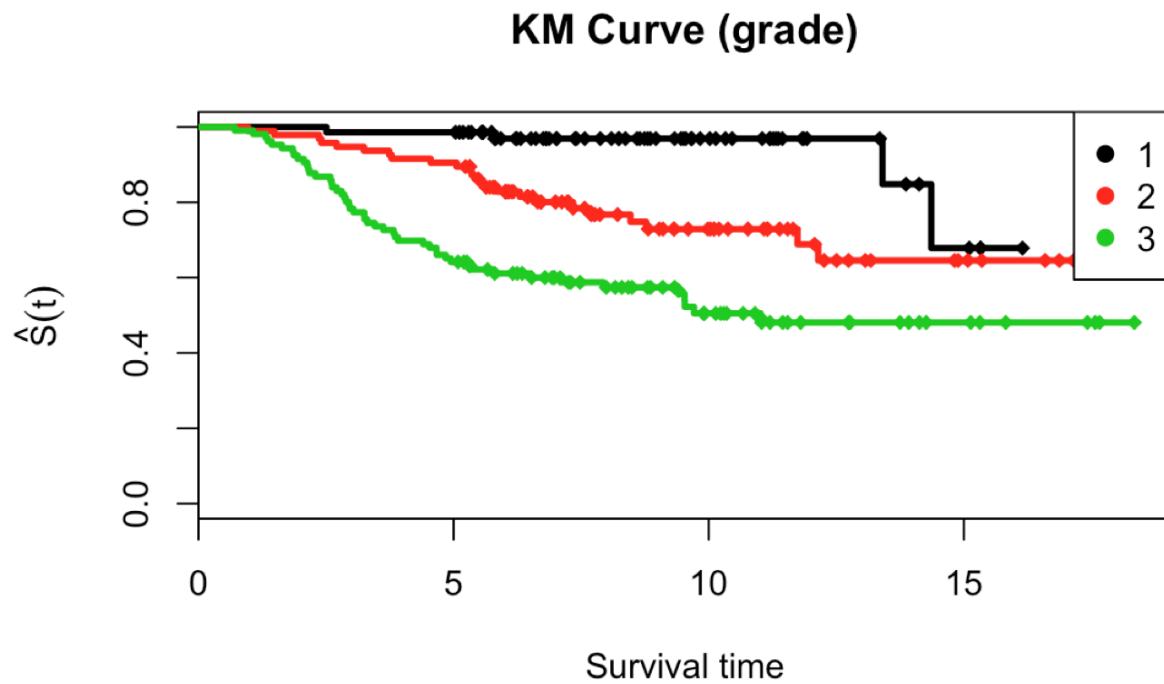


People who underwent amputation seemed to have a lower survival rate than people who didn't; but as mentioned, amputation is the last resort, so this may be because their condition was worse in the first place.

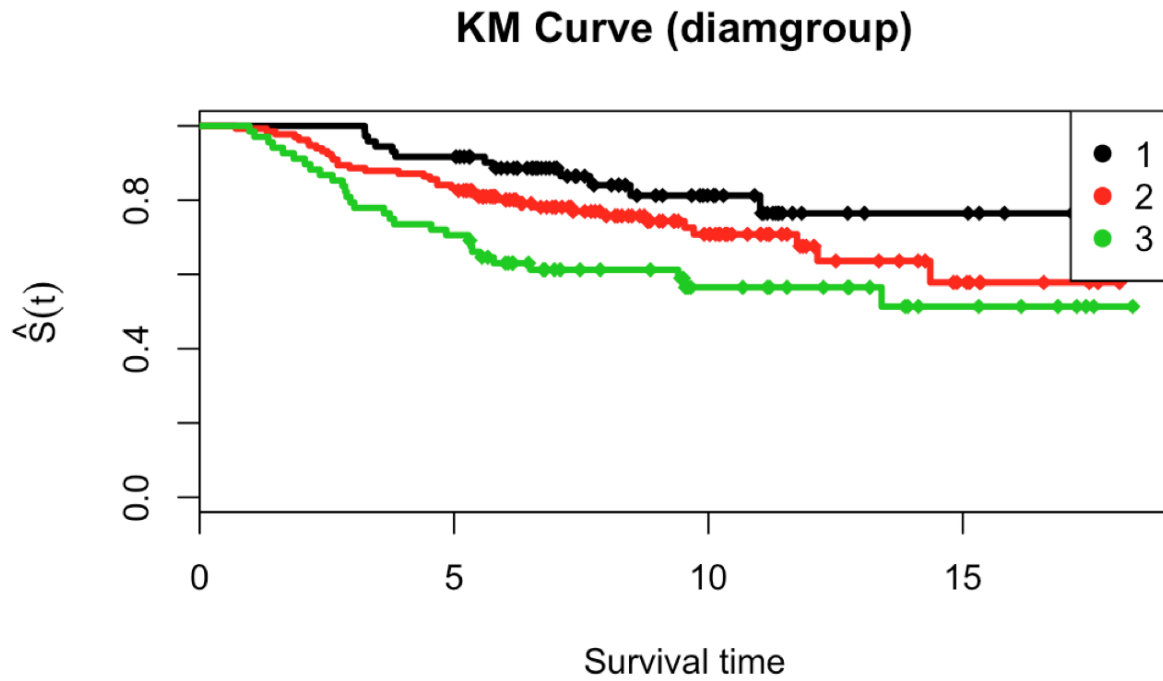


People who have hormonal therapy seem to have slightly higher survival rates than those who don't; but no recorded survival time for "hormonal" exists after about 14 years, while many people without hormonal therapy do survive after 14 years. Hormonal therapy may be used for worse conditions. Or maybe it has side effects on the body. We will need more domain expertise to interpret this.

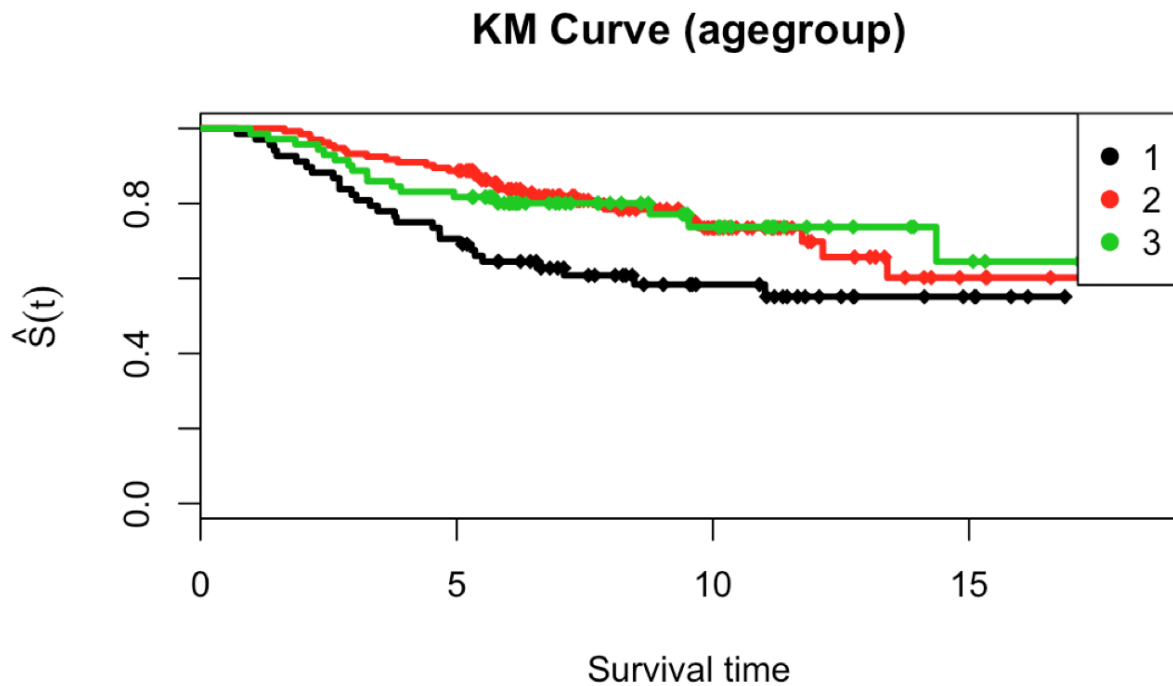
We then plot the three conditions: grade, diameter group, and age group.



Naturally, people with a higher grade of cancer are going to have shorter survival times in general, which corresponds to the graph: 1 at the top (highest rate of survival at all times), 2 in the middle, 3 in the bottom (lowest).



State '3' is the worst condition, and the plot confirms the “common sense” understanding. People with tumors which are larger than 30mm generally have the smallest survival rates at all times.



We may intuitively think that older people would have smaller survival rates. But here, it is actually the youngest group who has the least survival rates. And the oldest group, “3”, has lower rates than “2” in the beginning, but has similar rates between 6 and 12 years, and higher rates than “2” after about 12 years. Now, it is common knowledge that cancer may have something to do with the patient’s own body conditions. That is, if a person has a strong body (especially a strong immune system), they may be able to “fight off” some cancer cells, therefore be less likely to have cancer. So, the youngest patients may have a weaker immune system than their peers (the latter would be more likely to get cancer at an older age), which not only contributes to having cancer at a young age, but decreases the survival rates after they have cancer.

Log Rank Tests

After plotting Kaplan-Meier curves, we also conduct the Log-rank test on each variable. This would give a more rigorous test of the significance of each variable. However, not all of the p values are smaller than 0.05, which indicates not all variables have significant effects on the

event of death. From the R output, the p-value for the three covariates ‘grade’, ‘age’ and ‘diameter’ are less than 0,05, which means they all have significant effects on event death.

```
Call:
survdif(formula = NKI.surv ~ NKI$chemo)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
NKI\$chemo=0	165	51	46.8	0.378	0.964
NKI\$chemo=1	107	26	30.2	0.585	0.964

Chisq= 1 on 1 degrees of freedom, p= 0.3

```
Call:
survdif(formula = NKI.surv ~ NKI$hormonal)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
NKI\$hormonal=0	236	71	67.92	0.139	1.2
NKI\$hormonal=1	36	6	9.08	1.044	1.2

Chisq= 1.2 on 1 degrees of freedom, p= 0.3

```
Call:
survdif(formula = NKI.surv ~ NKI$amputation)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
NKI\$amputation=0	152	39	43.5	0.460	1.06
NKI\$amputation=1	120	38	33.5	0.596	1.06

Chisq= 1.1 on 1 degrees of freedom, p= 0.3

```
Call:
survdif(formula = NKI.surv ~ NKI$grade)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
NKI\$grade=1	71	4	22.3	15.032	21.20
NKI\$grade=2	95	24	28.4	0.672	1.07
NKI\$grade=3	106	49	26.3	19.549	29.79

Chisq= 35.4 on 2 degrees of freedom, p= 2e-08

```
Call:
survdif(formula = NKI.surv ~ NKI$agegroup)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
NKI\$agegroup=1	68	28	17.7	5.963	7.754
NKI\$agegroup=2	133	32	39.1	1.294	2.638
NKI\$agegroup=3	71	17	20.2	0.497	0.674

Chisq= 7.8 on 2 degrees of freedom, p= 0.02

```
Call:
survdif(formula = NKI.surv ~ NKI$diamgroup)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
NKI\$diamgroup=1	72	12	21.4	4.1479	5.774
NKI\$diamgroup=2	132	36	37.4	0.0531	0.103
NKI\$diamgroup=3	68	29	18.2	6.4657	8.519

Chisq= 10.8 on 2 degrees of freedom, p= 0.005

This confirms our observation from the Kaplan-Meier curves. The effects of the conditions are much more significant than treatments.

Model Building

After doing the log-rank test, we start to build our Cox PH model. We are using both the backward elimination method and the forward stepwise selection method to pick the right set of covariates. First, we build a full model with all covariates we chose from the log-rank step. Then we use the function “step” in R to apply the backward elimination method.

```
cox <- coxph(Surv(NKI$survival, NKI$eventdeath)~diamgroup+grade+agegroup, data = NKI)
step(cox, direction = "backward")
```

```

Start: AIC=768.56
Surv(NKI$survival, NKI$eventdeath) ~ diamgroup + grade + agegroup

              Df    AIC
<none>             768.56
- agegroup      1 770.02
- diamgroup     1 770.76
- grade         1 795.21

Call:
coxph(formula = Surv(NKI$survival, NKI$eventdeath) ~ diamgroup +
      grade + agegroup, data = NKI)

```

Then, we use the likelihood tests to select covariates. At the beginning we check the anova table for the model with the three significant covariates.

```

Analysis of Deviance Table
Cox model: response is Surv(NKI$survival, NKI$eventdeath)
Terms added sequentially (first to last)

      loglik   Chisq Df Pr(>|Chi|)
NULL      -403.69
diamgroup -398.42 10.5380  1  0.001169 **
grade     -383.01 30.8137  1  2.84e-08 ***
agegroup  -381.28  3.4623  1  0.062784 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

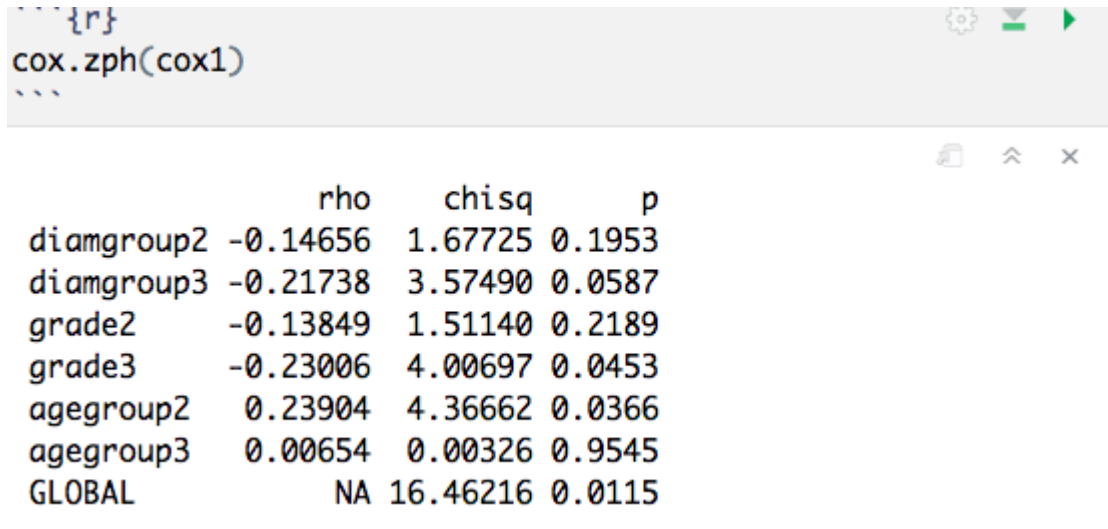
We choose the most significant variable ‘grade’ and then add ‘diameter’ to build a compared model. After running the likelihood test, we should include the ‘diameter’ covariate in our model. Similarly, we also add ‘age’ and conduct another likelihood test. In the end, we get the same model as using backward elimination method. But here, we see “agegroup” is not that significant (p-value 0.06), compared to “diamgroup” and “grade”, which both have p-values much smaller than 0.05. This confirms the finding from previous steps that “agegroup” and “survival” have a more complicated relationship.

Model Checking

We need to make sure all covariates meet the Cox PH assumption. We use both residual tests and C-log-log plots to check the Cox PH assumption. If the p-values and curves do not meet the assumption, we need to take a further step--stratify the covariate(s).

Residual tests

The function `cox.zph()` performs statistical tests on the PH assumption to test for independence between residuals and time. We find that the p-value of variable `grade2` and `agegroup2` are less than 0.05. It indicates the covariates `grade` and `agegroup` violates the PH assumption check. So we may need to stratify both `grade` and `agegroup`.



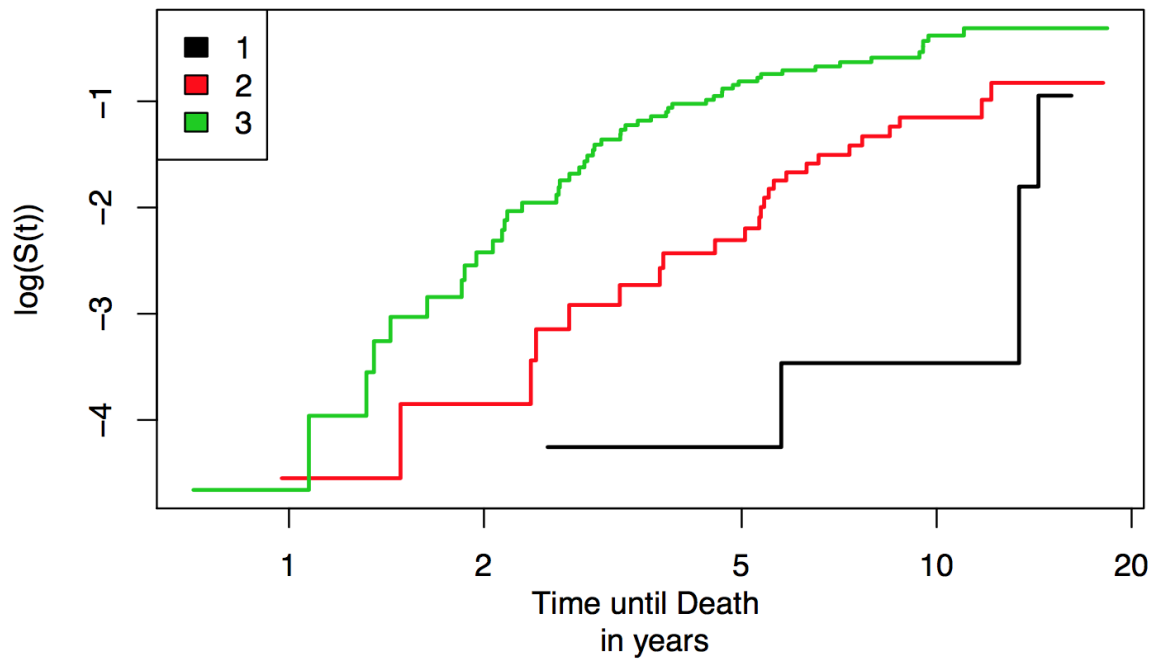
```
{r}  
cox.zph(cox1)  
...
```

	rho	chisq	p
diamgroup2	-0.14656	1.67725	0.1953
diamgroup3	-0.21738	3.57490	0.0587
grade2	-0.13849	1.51140	0.2189
grade3	-0.23006	4.00697	0.0453
agegroup2	0.23904	4.36662	0.0366
agegroup3	0.00654	0.00326	0.9545
GLOBAL	NA	16.46216	0.0115

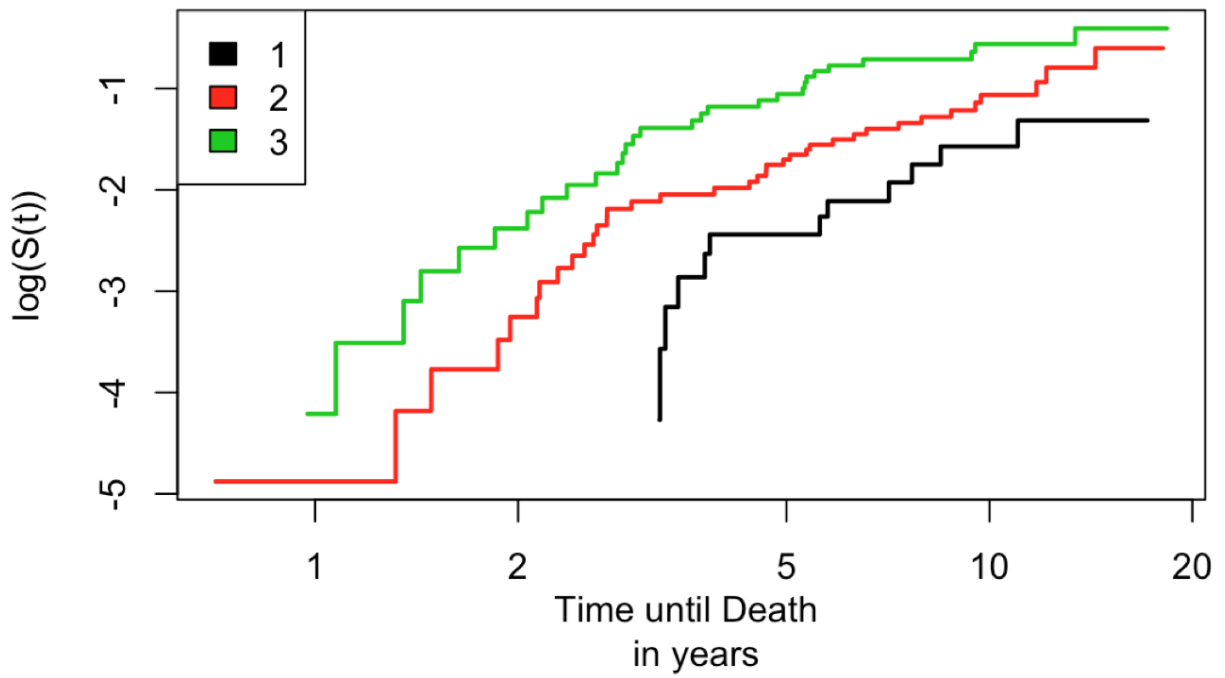
C-log-log plot

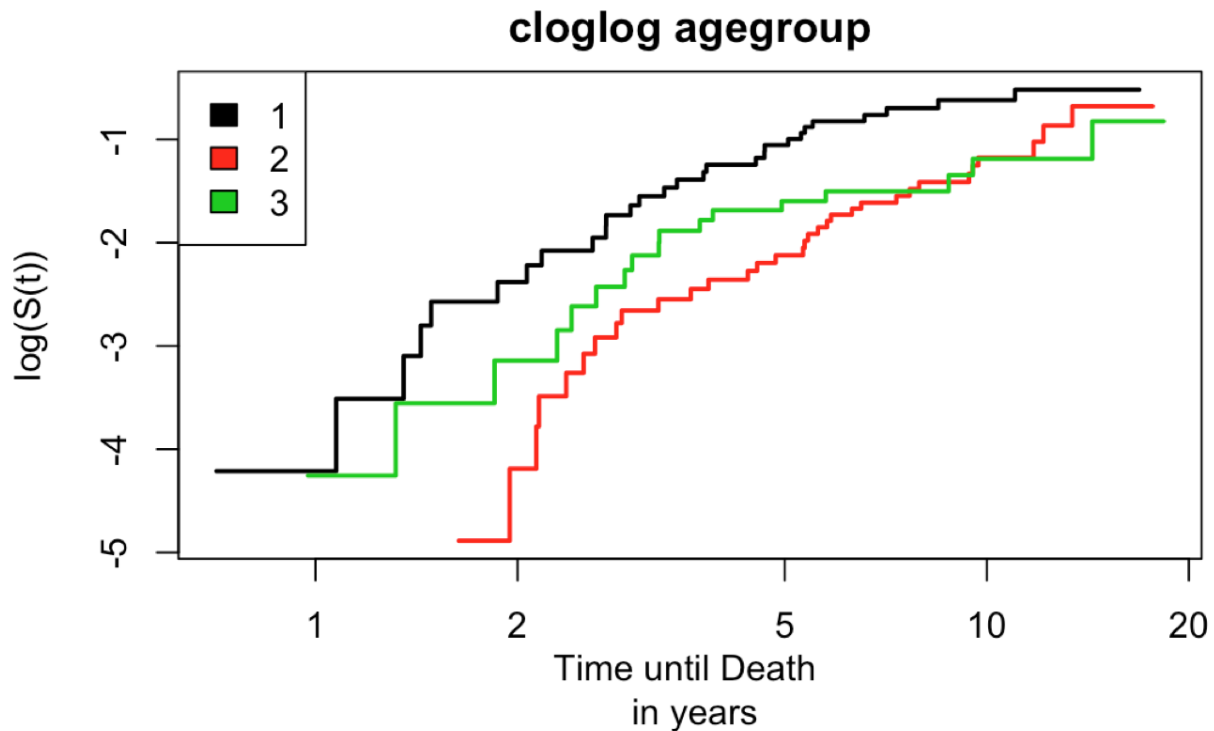
After performing the residual test, we check the C-log-log plot. From the C-log-log plot, we can see in a c log log plot, if two lines cross each other, that means that the covariate doesn't meet the Cox-PH assumption.

cloglog grade



cloglog diamgroup





In the “residual tests” step, both “grade” and “agegroup” have p-values smaller than 0.05 (don’t meet the coxPH assumption), and we can also see that the “agegroup” and “grade” both have a crossing problem. So, we decide to stratify on “agegroup” and “grade”.

Interaction term

After performing these tests regarding the individual covariates, we start to consider whether interaction terms will have a significant effect on our model. There are 3 potential interaction terms: `strata(agegroup)*diamgroup`, `strata(agegroup)*strata(grade)`, `diamgroup*strata(grade)`. We check the anova table for all three models with interaction terms.

Analysis of Deviance Table

Cox model: response is Surv(NKI\$survival, NKI\$eventdeath)
Terms added sequentially (first to last)

	loglik	Chisq	Df	Pr(> Chi)
NULL	-318.38			
diamgroup	-312.27	12.208	2	0.002234 **
diamgroup:strata(agegroup)	-310.86	2.813	4	0.589594

Cox model: response is Surv(NKI\$survival, NKI\$eventdeath)
Terms added sequentially (first to last)

	loglik	Chisq	Df	Pr(> Chi)
NULL	-322.71			
diamgroup	-320.63	4.1711	2	0.1242
diamgroup:strata(grade)	-320.43	0.3944	4	0.9829

Cox model: response is Surv(NKI\$survival, NKI\$eventdeath)
Terms added sequentially (first to last)

	loglik	Chisq	Df	Pr(> Chi)
NULL	-240.06			
strata(agegroup):strata(grade)	-240.06	0	3	1

All the interaction terms have large p-values (much greater than 0.05). Therefore, we will not include these interaction terms in our model. Therefore, our model is $\text{Surv} \sim \text{strata}(\text{agegroup}) + \text{diamgroup} + \text{strata}(\text{grade})$.

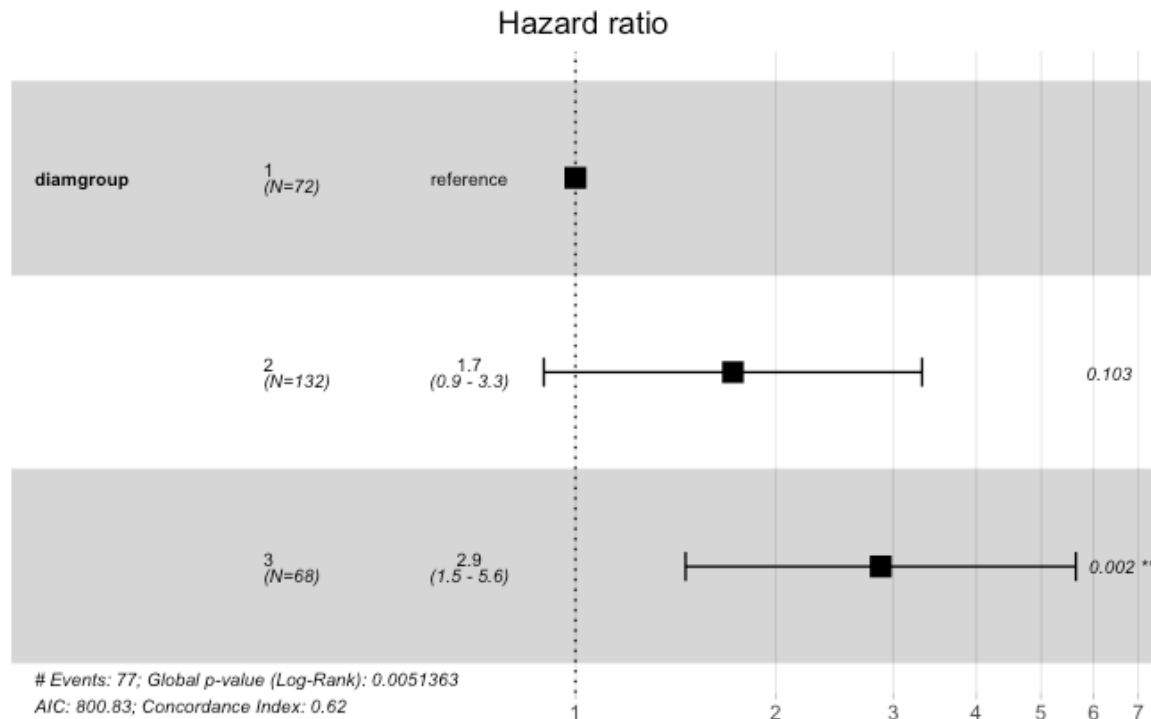
We do another `cox.zph` test for this final model. This model satisfies the `cox.zph` assumptions.

```
```{r cox_zph|stratify}
cox2 <- coxph(Surv(NKI$survival,
NKI$eventdeath)~diamgroup+strata(grade)+strata(agegroup), data = NKI)
cox.zph(cox2)
```
```

| | rho | chisq | p |
|------------|--------|-------|--------|
| diamgroup2 | -0.166 | 2.18 | 0.1396 |
| diamgroup3 | -0.205 | 3.24 | 0.0719 |
| GLOBAL | NA | 3.30 | 0.1918 |

Hazard Ratios and C.I.

We are using `ggforest()` to visually showing the Hazard Ratios and Confidence Intervals for diameters of different groups.

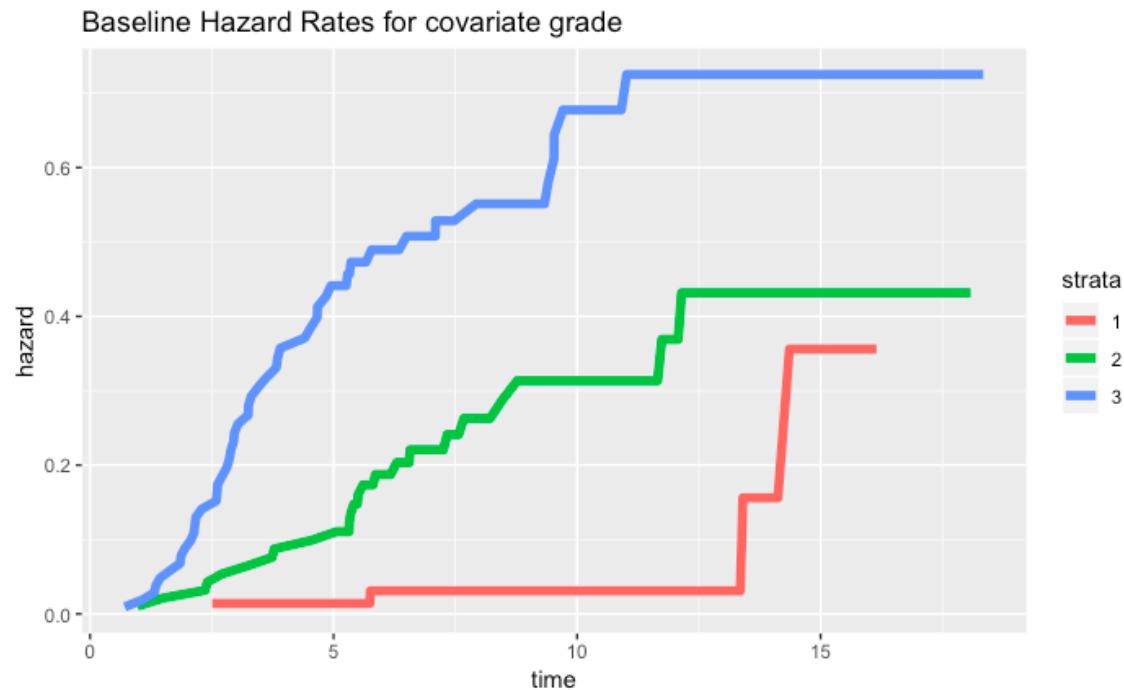


From the hazard ratio chart above, we know that the hazard ratio for diamgorup 2 is centered at 1.7 and its 95% confidence interval is between 0.9 and 3.3. It indicates that patients' tumors which are larger than 15 but less than 30 have 70% more likelihood to die than patients' less than 15mm(group1). The hazard ratio for diamgorup 3 is centered at 2.9 and its 95% confidence interval is between 1.5 and 5.6. It indicates that patients' tumors which are larger than 30 mm have 190% more likelihood to die than patients' less than 15mm(group1).

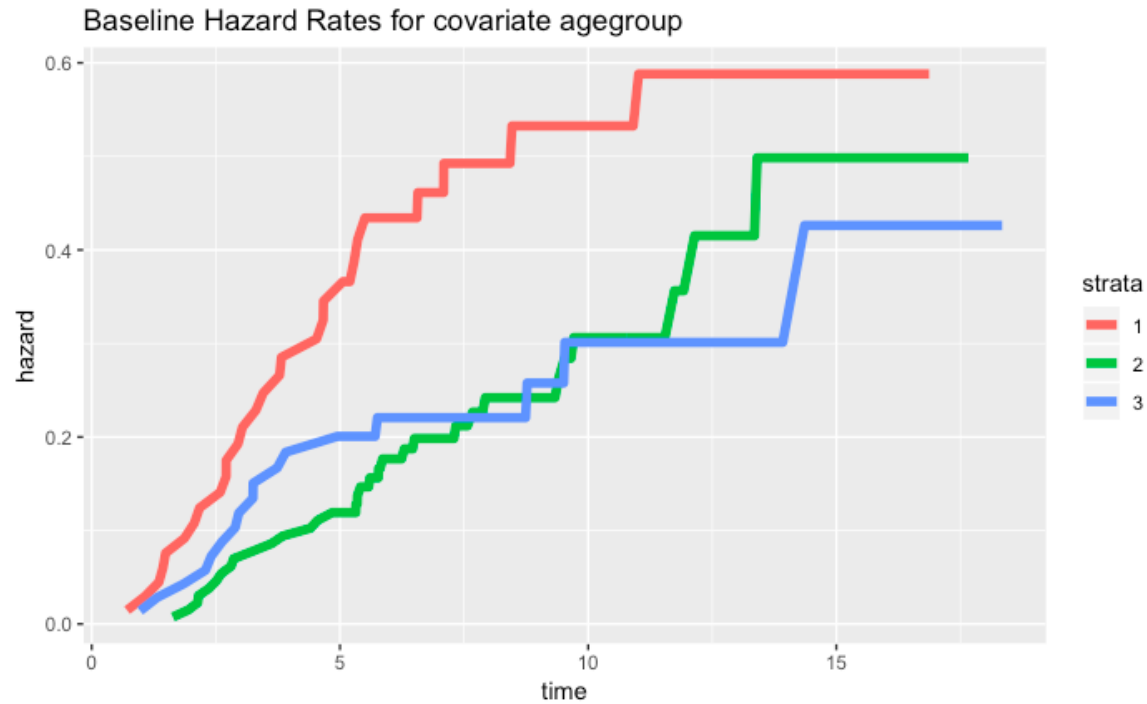
Baseline Hazard Rates

We also draw the baseline hazard plot for each strata (1 indicates grade level 1 , 2 indicates grade level 2, 3 indicates grade level 3). We can get the baseline hazards ($h_0(t)$) at

different lengths for each group. From the plot below, we clearly see that grade level 3 has higher baseline hazard rates than grade level 2 and grade level 1. As a result, level 3 (patients with the fastest growing cancer) are more likely to die than level 1 and level 2. The faster the cancer grows, and the more “abnormal” the cells look like, the more likely it is for the patient to die.

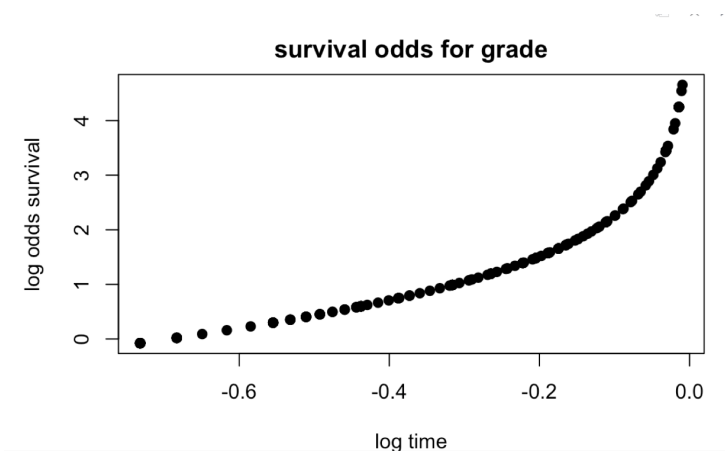


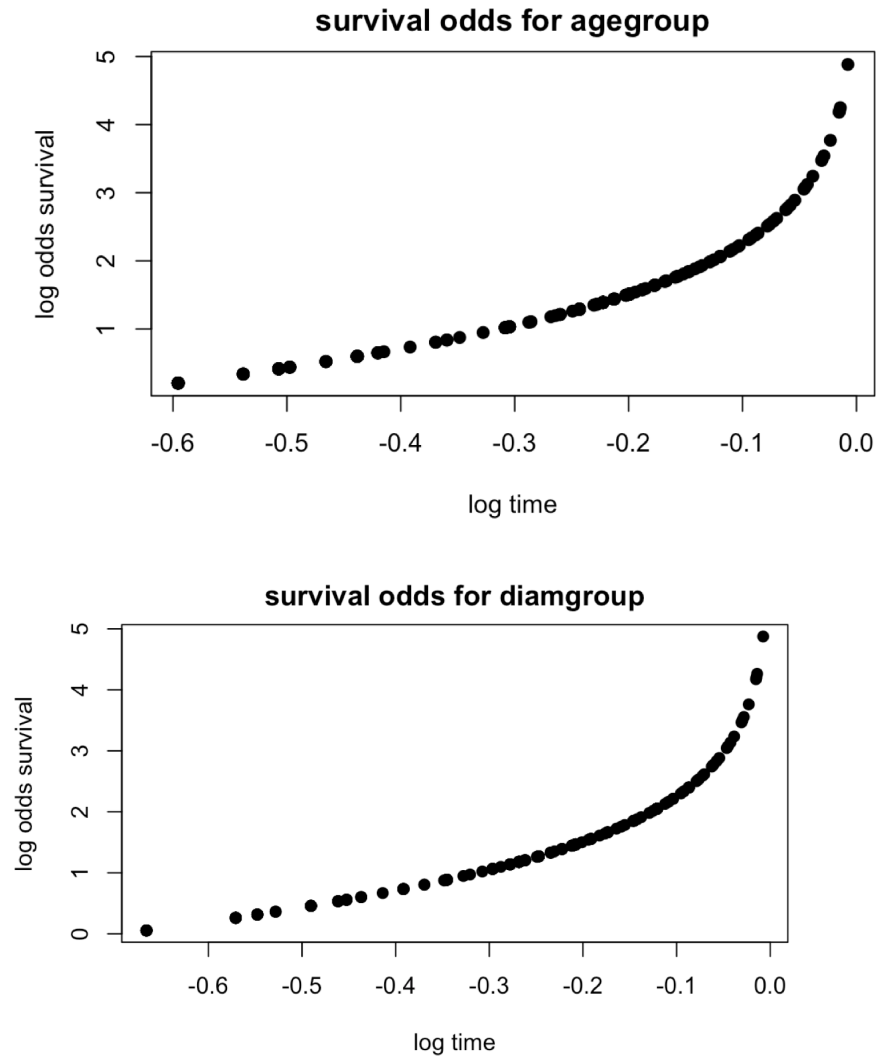
The baseline hazard rates for agegroup are more complicated. Agegroup 1 has a consistently greater hazard rate. Before 7.5 years, agegroup 3 hazard rate is higher than agegroup 2; from 7.5 to 12, similar; after 12, agegroup 3 becomes the lowest. So, age does not have a simple relationship with hazard rate.



Extension-AFT

We try to build AFT model with log-logistic distribution, so we check our assumption by drawing the survival odds plot for variables grade, diamgroup, and agegroup. Straight lines indicate a log-logistic distribution. These don't have straight lines, so we can't use the log-logistic distribution.





Conclusion

In this project, we were given a right-censored data of 272 patients along with the covariates that illustrates their conditions and treatments: diameter of tumors, age and their grade levels of cancer; treatments are chemotherapy, hormonal therapy, and amputation. We plot Kaplan-Meier curves for each covariate, and use the log-rank tests, to check if any of the covariates affect survival time. Both the Kaplan-Meier curves and the log-rank test indicate that the three conditions have significant effect on survival time, while the treatments don't have a significant effect.

We proceed to build a model using the backward elimination method and the forward stepwise method. The model built included only the three conditions: age group, diameter group, and grade of cancer. While age group's relationship with survival time is a little complicated, we include it in the initial model.

Next, we do residual tests (cox.zph) and the c-log-log tests. The cox.zph and C-log-log plots indicate that the initial model does not meet the proportional hazards assumption, which resulted in stratification of variables (grade) and (agegroup) in our model. We then tested for possible interaction terms, and concluded that none of them should be included in the model. Since none of the interaction terms have p-values smaller than 0.05, we do not include any in our model.

Finally, we estimated the hazard ratios and confidence intervals of diameter group (not stratified), and baseline hazard rates for the grade and age group (stratified). We found that grade of breast cancers is clearly related to survival time. The higher grades' cancers are more likely causing death because they grow faster and do more harm to humans. When we focus on age, the relationship is more complex. Before 7.5 years, agegroup3 has a higher hazard rate than agegroup2. After 12 years, the agegroup2 shows a higher hazard rate than agegroup3. We concluded that age is not simply related to hazard rate.

As an extension, we also attempted the AFT model with log-logistic distribution. But the survival odds plot doesn't indicate such distribution, so we can't use it.

Overall, we do have some clear takeaways from this research. First, the effects of single treatments are not significant to survival time. Second, survival time depends more on the condition of the patient and the tumors. However, that is not to say the treatments aren't effective. Third, counter-intuitively, age's relationship with survival time isn't straightforward. We believe this project serves as a start, where we did survival analysis on six of the many possible covariates that are collected on a breast cancer patient.

References

- (1) Size and grade of breast cancer: <https://breastcancernow.org/information-support/facing-breast-cancer/diagnosed-breast-cancer/cancer-grade-size>
- (2) Explanation of certain variables, like "amputation": https://file.scirp.org/pdf/JBiSE_2018052914441283.pdf
- (3) Breast Cancer Stages: <https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html>