German Breast Cancer Survival Analysis

-Udyot Kumar

CONTENTS

List of Tables

List of Figures

A) BACKGROUND

The following data set has been collected from German breast cancer study conducted in 1989. Such Cancer clinical trials provide a good source for time to event analysis. The study contains 686 observations with 16 variables. Out of 686 patients 299 had a recurrence of cancer within the study interval while 171 died.

In this project, I have used Survival Analysis algorithm which is designed for analysis of duration data i.e. How long until an event occurs. In our case, it is how long will a patient survive.

Survival analysis involves a time to event or survival time variable, in our case it is number of days and status of 'event of interest' i.e. death/alive or recurrence/no recurrence. One thing to note is that all patients do not die in survival analysis so Survival time is not available for all the patients. Survival time for such patients is considered as the duration of study. From this, an important concept of censoring comes into the picture. Censoring means we are not able to collect information about the patients who didn't die during the time duration we studied. And that's the main reason we prefer survival analysis over Logistic regression. Problem with Logistic Regression is that it ignores censoring and provide skewed output as it doesn't consider how long the patient survived.

- Assumptions of Survival Analysis:
    - Participants are independent
    - Events must represent change of state from one form to another
    - Participants are free when they enter the study
    - It assumes time data is continuous
    - Censoring is unrelated to probability of event occurrence

- Two things are recorded for each participant:
    - The know time for each participant in which reoccurrence didn't happen
    - Status of participant on the last day he/she was observed

We will also evaluate the covariates that are predictive of time to death and breast cancer recurrence. The eight covariates used in this study are- age, menopause status, grade of cancer cells, tumor size, number of positive lymph nodes, hormonal therapy, progesterone and estrogen receptors concentration.

In survival analysis, the two key functions are survival function and Hazard function. Survival function gives the probability of surviving till a specified time. The hazard function gives the potential that event will occur given that patient survived up to a specified time.

A semi parametric model which we will use in our survival analysis will be Cox proportional hazard regression model. It is the most widely used method. A non-parametric estimator of our survival analysis will be Kaplan Meier. And for Parametric model we will use exponential distribution, Weibull and Log-normal.

B) INTRODUCTION

Analysis of treatment and various factors which affect the cancer is a popular topic of research. One of the various factors in this study is grade of cancer cells which is defined by rate of growth of cancer cells. Grade 1 is slowest growing cells and they resemble the normal cells, grade 2 grows faster than grade 1 while grade 3 are abnormal cells and grows most rapidly. Cancer cells sometimes travel through bloodstream and reach lymph nodes.
The number of lymph nodes involved with cancer cells in our study varies from 1- 51 and is probably one of the most important factors to study.
Progesterone receptors (PR) is a protein found in cells activated by hormone progesterone. Estrogen receptors (ER) are activated by estrogen hormone. PR count ranges from 1-2380 and ER from 1-1144.
Tumor varies in shape and size. Tumor size is measured in millimeters and they most commonly range from 10 mm to 50 mm in diameter. In our study the min size is 3 mm while max is 120 mm.

Menopause status and Hormonal therapy are two categorical variables. 440 patients were going through Hormonal therapy while 290 patient's menopause status was yes.

C) DATA DESCRIPTION

The data set has 686 observations and 16 variables. The list of variables is:

Variable	Name	Description	Codes/Values
1	id	Study ID	1 - 686
2	diagdate	Date of Diagnosis	ddmmmyyyy
3	recdate	Date of Recurrence or of Recurrence Free Survival	ddmmmyyyy
4	deathdate	Date of Death	ddmmmyyyy
5	age	Age at Diagnosis	Years
6	menopause	Menopausal Status	1 = Yes, 2 = No
7	hormone	Hormone Therapy	1 = Yes, 2 = No
8	size	Tumor Size	mm
9	grade	Tumor Grade	1 - 3
10	nodes	Number of Nodes involved	1 - 51
11	prog_recp	Number of Progesterone Receptors	1 - 2380
12	estrg_recp	Number of Estrogen Receptors	1 - 1144
13	rectime	Time to Recurrence	Days
14	censrec	Recurrence Censoring	0 = Censored; 1 = Recurrence
15	survtime	Time to Death	Days
16	censdead	Death Censoring	0 = Censored; 1= Death

Table 1

D) DISTRIBUTIONS

Summary statistics for continuous variables:

Age	
Min.	21
1st quarter	46
Median	53
Mean	53.05
3rd quarter	61
Max.	80

size	
Min.	3
1st quarter	20
Median	25
Mean	29.33
3rd quarter	35
Max.	120

nodes	
Min.	1
1st quarter	1
Median	3
Mean	5.01
3rd quarter	7
Max.	2380

prog_recp	
Min.	0
1st quarter	7
Median	32.5
Mean	110
3rd quarter	131.8
Max.	2380

estrg_recp	
Min.	0
1st quarter	8
Median	36
Mean	96.25
3rd quarter	114
Max.	1144

rectime	
Min.	8
1st quarter	567.8
Median	1084
Mean	1124.5
3rd quarter	1684.8
Max.	2659

Table 2

D1) Distribution of age:

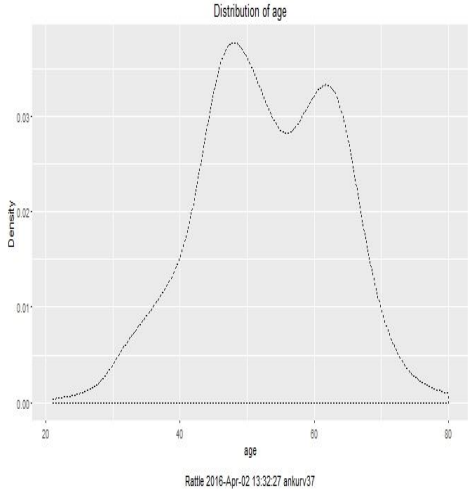


Figure 1: Age

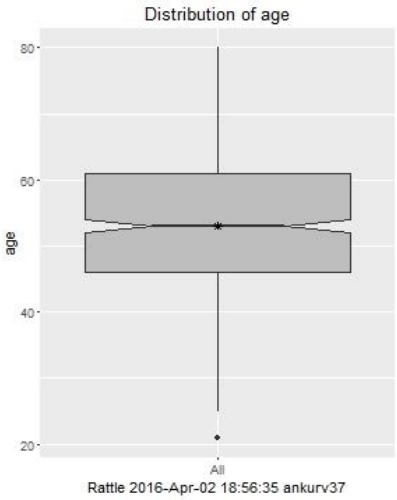


Figure 2: box plot- age

As we can see from the above plots that the age distribution of the subjects are largely in between 45-65.

D2) Distribution of lymph nodes:

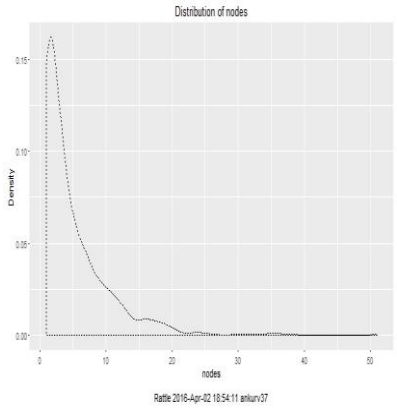


Figure 3: Lymph nodes

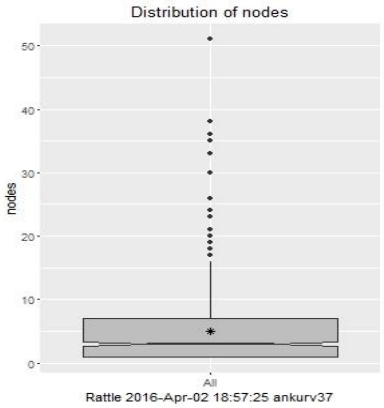


Figure 4: Box plot

Well we can see from the plots that mean number of lymph nodes (number of swellings in the lymphatic system where lymph is filtered and lymphocytes are formed) is around 5.

Just for information:

- Lymph node-negative means the axillary lymph nodes do not contain cancer.
- Lymph node-positive means the axillary lymph nodes contain cancer.

  - See more at: http://ww5.komen.org/BreastCancer/LymphNodeStatus.html#sthash.vpvVyFhh.dpuf

D3) Distribution of Progesterone receptors:

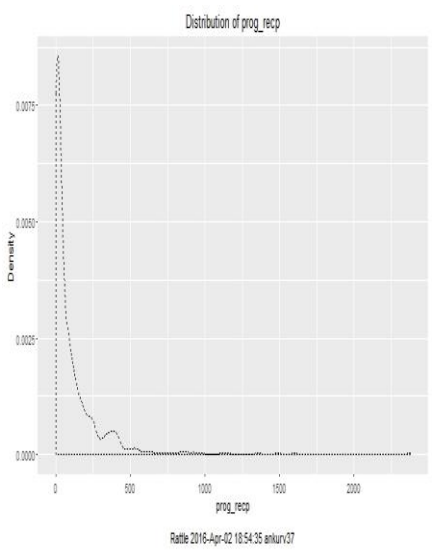


Figure 5: Progesterone hormone

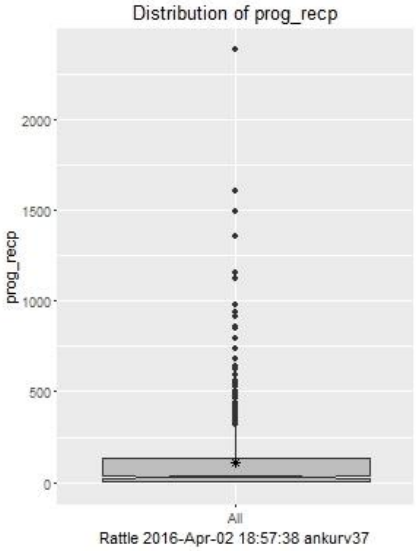


Figure 6: Box plot

The mean number of progesterone receptors are 110. Outliers are present as we can see from box plot.

D5) Distribution of Estrogen receptors :

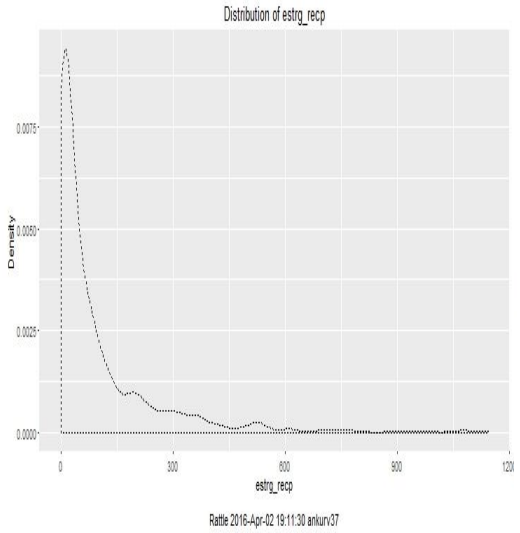


Figure 7: Estrogen hormones

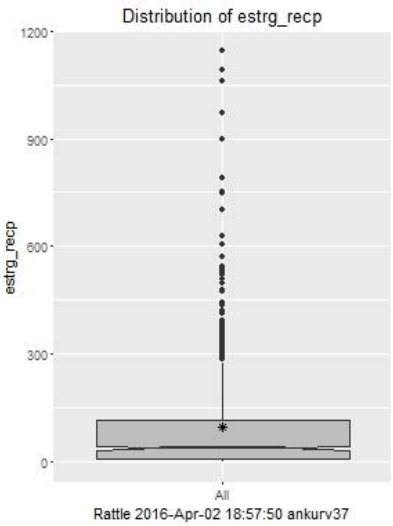


Figure 8: Box plot

The mean number of estrogen receptors are 96.25

Key Points:

- Estrogen and progesterone receptors are found in breast cancer cells that depend on estrogen and related hormones to grow.
- All patients with invasive breast cancer or a breast cancer recurrence should have their tumors tested for estrogen and progesterone receptors.

D6) Distribution of Tumor size:

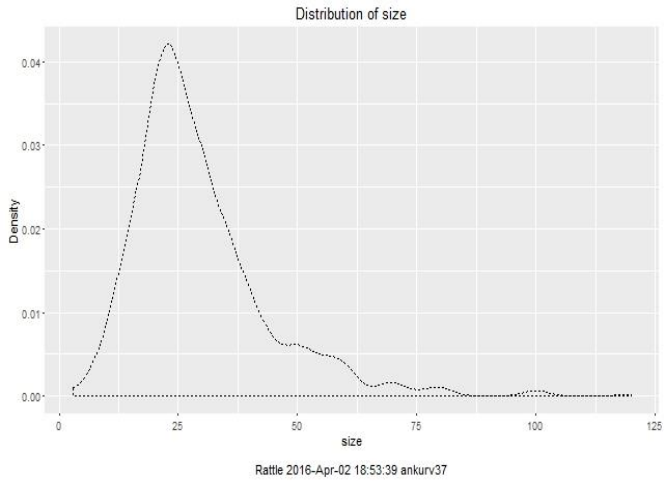


Figure 9: Tumor size

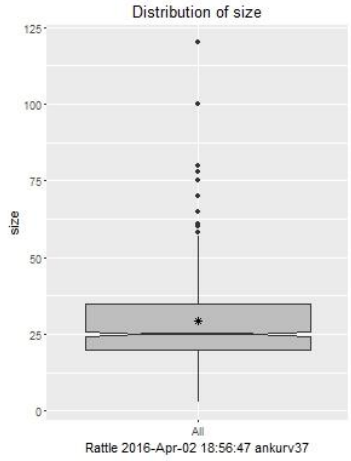


Figure 10:Box plot

The mean is 29.33

E) CORRELATION PLOT:

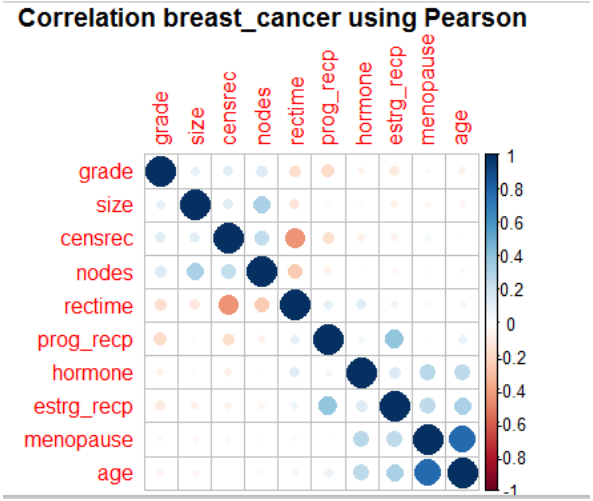


Figure 11: Correlation

There is high correlation between menopause and age which is correct. Estrogen and Progesterone level are positively correlated too.

F) Kaplan-Meier Analysis:

Next, I ran some basic Kaplan-Meier plots to further visualize the data, as well as see which factors can play a role in determining time to death. The Kaplan-Meier estimator can be defined as the non-parametric estimator of the survival function. The KM plot shows time to death on the x-axis, with the percentage of patients still surviving on the y-axis. Below is a basic KM plot for our Breast Cancer data:

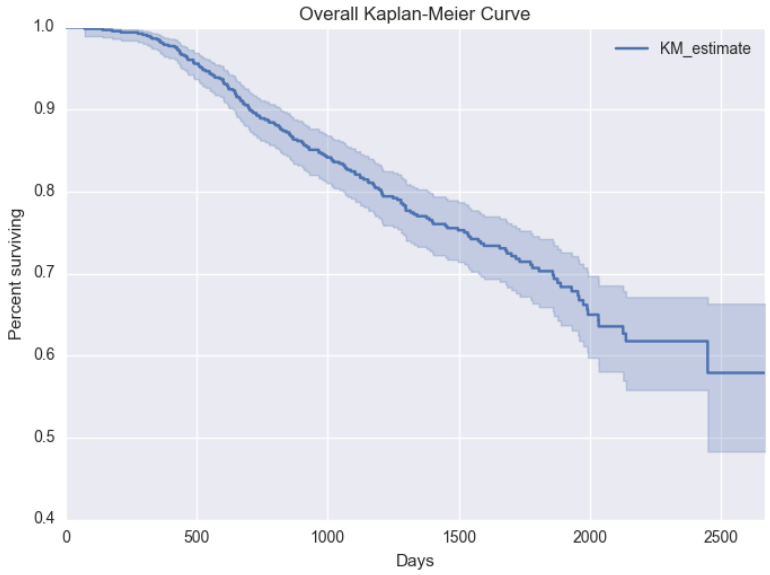


Figure 12: Survival rate- KM

So, each time you see the line above drop down, a patient has died. You can see that the line ends at about a 58% survival rate after about 2,600 days. The survival function cannot rise over time because once someone relapses it cannot be undone.

How to read it?

- Patients who have gone through therapy have 0.84 probability of surviving for at least 1000 days.
- Patients who have gone through therapy have 0.58 probability of surviving for at least 2500 days.

Overall, the KM analysis was a good starting point to visualize the survival function.

G) Cox Proportional Hazard Analysis:

Next up, we ran some Cox Proportional Hazard analysis. Cox PH analysis is the standard and most popular use in Survival Analysis. When running a Cox PH model, we are looking at the

relationships between the survival of a patient and various explanatory variables. We are looking to understand whether specific variables increase, decrease, or have no effect on the probability of our death event.

When running a Cox PH model, the output gives each explanatory variable a hazard ratio coefficient. If this coefficient is equal to one, it means that the variable has no effect on the death event. If the coefficient is greater than one, it means the variable increases the hazard, or the likelihood of the death event. Finally, if the coefficient is less than one, it means the variable reduces hazard, or the likelihood of the death event.

Assumptions of COX PH Model:

- The Cox regression model is a semiparametric model, making fewer assumptions than typical parametric methods but more assumptions than those nonparametric methods.
- It makes no assumptions about the shape of the so-called baseline hazard function.
- H(t)= H0(t)exp(A+Bx) and it yields hazard rate
- While a nonlinear relationship between the hazard function and the predictors is assumed
- The hazard ratio comparing any two observations is in fact constant over time in the setting where the predictor variables do not vary over time. This assumption is called the proportional hazards assumption.
- It allows testing for differences in survival times of two or more groups of interest, while allowing to adjust for covariates of interest.
- Kaplan Meier curves helps us in validating the output given by cox PH

Before building the model, I also checked if all the variables are following COX PH assumption. It is optional and you can find the report on that after the model.

Model built in R is below:

Model 1

- *coxph(formula = Surv(survtime, censdead) ~ age + menopause + hormone + size + grade +nodes + prog_recp + estrg_recp, data = crs$dataset[, c(crs$input, crs$risk, crs$target)])*
- n= 686, number of events= 171

covariates	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
age	0.006702	1.006725	0.01212	0.553	0.58026	
menopause	0.100006	1.105178	0.25265	0.396	0.69223	
hormone	-0.26516	0.767082	0.168608	-1.573	0.1158	
size	0.013244	1.013332	0.00483	2.742	0.0061	**
grade	0.42748	1.533388	0.145752	2.933	0.00336	**
nodes	0.052942	1.054368	0.00953	5.556	2.77E-08	***
prog_recp	-0.00535	0.994667	0.001184	-4.515	6.3414E-06	***
estrg_recp	-0.00028	0.999716	0.000547	-0.519	0.60394	

Table 3: Cox-ph results model 1

The exp(coef) variable is the Hazard ratio:

- HR = 1 : No effect
- HR < 1 : Reduction in hazard (Death)
- HR > 1 : Increase in Hazard (Death)
- For explanatory variables that are continuous (for example, age) the regression coefficient refers to the increase in log hazard for an increase of 1 in the value of the covariate
- The estimated hazard in the hormone=yes group is exp(-0.26516) = 0.767082 of that of the total group; that is, a decrease in the risk of death after adjustment for the other

explanatory variables in the model. However, p-value is not statistically significant suggesting it will not affect survival.

Model 2

In the second model, I stratified the age which was a continuous variable into three groups- young (age =< 30), mid age (30-50), old age (50>=). The aim was to observe if the age group as a categorical variable could give us any insight different than previous model and if they are significant.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
age2	-1.3674115	0.2547656	0.5269436	-2.595	0.00946	**
age3	-0.7930698	0.4524537	0.5690171	-1.394	0.16339	
menopause2	-0.2094036	0.8110678	0.2682731	-0.781	0.43506	
hormone	-0.2600882	0.7709836	0.1696605	-1.533	0.12528	
size	0.0148067	1.0149168	0.0047867	3.093	0.00198	**
grade2	0.7383108	2.0923981	0.4254175	1.735	0.08265	.
grade3	1.0890069	2.9713217	0.4411735	2.468	0.01357	*
nodes	0.0527359	1.0541512	0.0095067	5.547	2.90E-08	***
prog_recp	-0.0053014	0.9947126	0.0011814	-4.487	7.21E-06	***
estrg_recp	-0.0002726	0.9997274	0.0005405	-0.504	0.61402	

Table 4: Cox-ph model 2

- We find that age2 which is age range from 30-50 has a hazard ratio of 0.254 which proves that this group has a lower risk of hazard when after adjustment for other explanatory variables.
- We also observe that grade of cancer cells was also a significant factor in the first model itself, but now we recognize that grade 3- the fastest growing cancer cells which looks

abnormal in contrast with normal cells have a much higher hazard ratio (2.97) when compared to other groups, also the p-value is statistically significant in this case.

Final Model

- `coxph(formula = Surv(time, event) ~ age + size + grade + prog_recp + nodes, method = "breslow")`
- `n= 686, number of events= 171`

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
age2	-1.450841	0.234373	0.525153	-2.763	0.00573	**
age3	-1.111917	0.328928	0.514399	-2.162	0.03065	*
Size	0.01536	1.015479	0.004731	3.246	0.00117	**
grade2	0.753265	2.123924	0.424961	1.773	0.0763	.
grade3	1.13428	3.108935	0.440413	2.575	0.01001	*
prog_recp	-0.005297	0.994717	0.001148	-4.614	3.95E-06	***
Nodes	0.051101	1.05243	0.009422	5.423	5.85E-08	***

Table 5: Cox-ph results – Final model - R

- So, we arrive at conclusion that Age, size, grade, progesterone hormone receptors and number of positive lymph nodes explain the survival time.
- Number of positive lymph nodes to where cancer has spread is a significant factor which increases the hazard (or death) by a factor of 1.05.
- Number of progesterone hormone receptors is another significant but one which reduces hazard by 0.99.
- Grade 3 cancer cells are more significant than grade 2, which makes sense as grade 3 grow faster than grade 2, but overall they both increase the hazard by biggest margin of > 2.
- Middle age group(30-50) have a lower hazard than patients with older age by a big margin of 0.22.

OPTIONAL

I also tried to test the assumptions of Cox PH method. It's optional but a good thing to learn and understand. It's as follows:

Validating Proportional Hazard assumption:

a) Using Kaplan Meier Plots

- We will now break this plot down by different factors. The categorical variable will split the main line into multiple sub categories. COX PH is an important assumption which assumes that for time independent variables the hazard functions must be proportional over time. So to prove this we create Kaplan Meier Curve for Tumor Grade Category and split out by Tumor Grade (1-3):

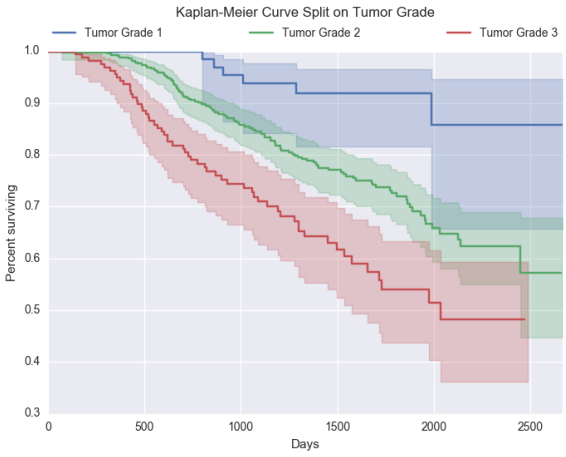


Figure 13: Survival rate- Tumor grade

You can see that these lines don't ever cross, and they have constant distance between them. This proves the assumption of proportional hazard for tumor grade as percentage of surviving for each tumor grade remains proportional over time.

We also see that those with a tumor grade of 1 have a much higher probability of surviving over the lifetime of the observation, as compared to tumor grade 2 or 3.

- Here is the Kaplan-Meier Curve split out by whether the patient has had menopause or not.

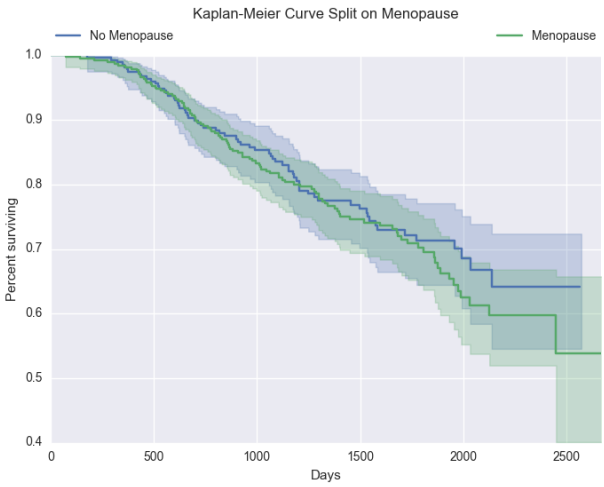


Figure 14: Survival rate - Menopause

You can see that these lines cross back and forth many times, violating the proportional hazard assumption. But when we check our COX result we notice that Menopause is not a significant variable and that is why it is not included in our final model.

- We also ran a KM curve, splitting it by whether the patient had received hormone therapy or not.

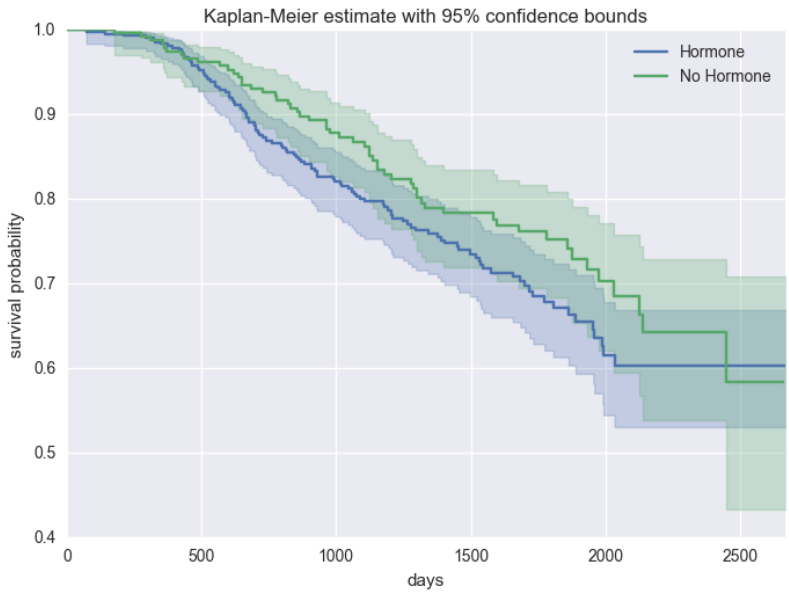


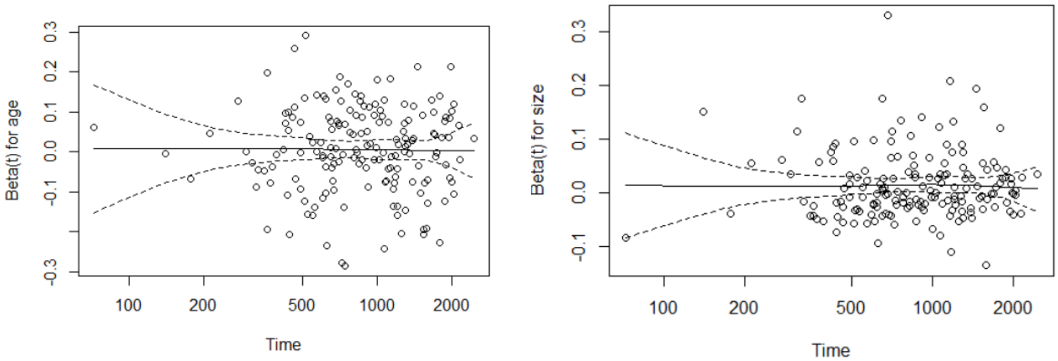
Figure 15: Survival plot - Hormones

You can see that, for the most part, the no-hormone group had a higher survival chance until the very end, where the two lines ended up crossing. But after finding the final model above we noticed that Hormone therapy is not included in the final result as it is not significant.

b) Tests and Graphs based on Schoenfeld Residuals:

In this test we check the Proportionality Hazard assumption for time dependent covariates. Testing the time dependent covariates is equivalent to testing for a non-zero slope in a generalized linear regression of the scaled Schoenfeld residuals on functions of time. A non-zero slope is an indication of a violation of the proportional hazard assumption.

Here first we have created coxph object by using coxph function in R. To create the plots of the Schoenfeld residuals versus log(time) create a cox.zph object by applying the **cox.zph** function to the cox.ph object. Then the **plot** function will automatically create the Schoenfeld residual plots for each of the predictors in the model including a lowess smoothing curve.



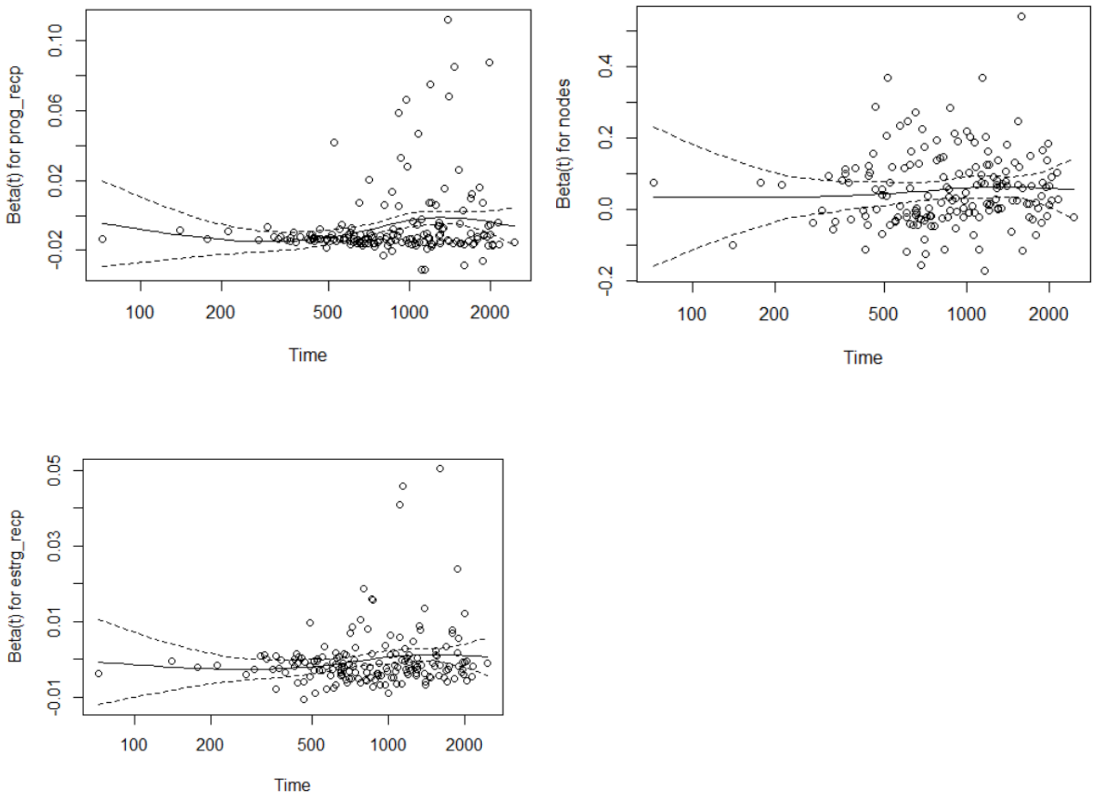


Figure 16: Schoenfeld Residuals vs Time

We see in the above graphs that the residuals are not having any slope. Though in Progesterone and Estrogen plots slope is not exactly 0 but we cannot find any pattern by looking at the residuals so COX PH assumption is not violated.

H) Parametric Regression Models

Finally, I ran a few different parametric models on the data. I used an exponential, Weibull, and log-normal distribution. I ran the regression using different categorical variables, like I did above for the KM curves.

First, ran the regression models for the various tumor grades:

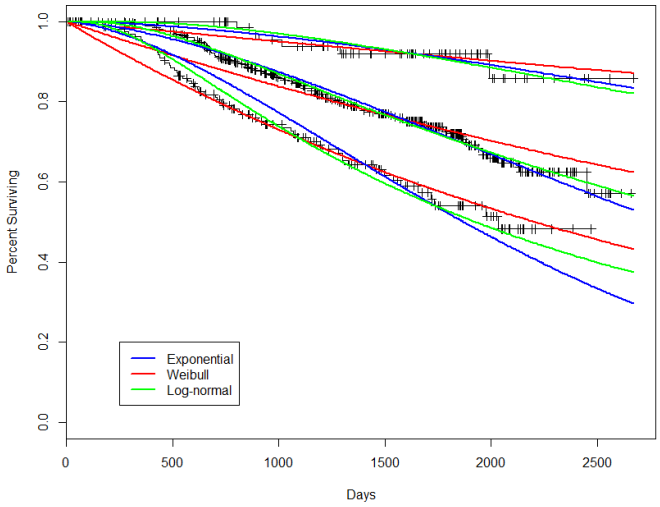


Figure 17: Parametric model- Tumor grade

Tumor Grade			
	Exponential	Weibull	Log-normal
AIC	3,251.22	3,213.84	3,197.75

Figure 18: Parametric model comparison using AIC – Tumor grade

By looking at the AIC values for the different models above, we can see that the Log-normal model performed the best, albeit by only a very slim margin. On the plot, you can see each of these regression lines laid over the KM curve for tumor grades 1, 2 and 3.

We also used the same approach towards the Hormone variable.

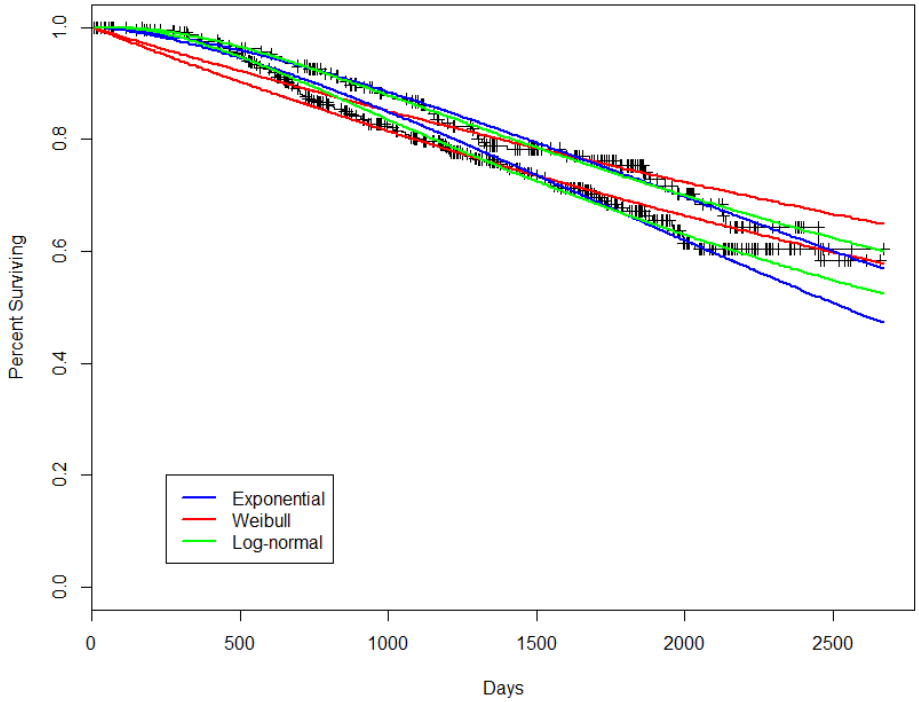


Figure 19: Parametric model- Hormone therapy

Hormone Therapy			
	Exponential	Weibull	Log-normal
AIC	3,276.52	3,242.43	3,229.67

Figure 20: Parametric model comparison using AIC – Hormone therapy

Again, the log-normal model provided us with the lowest AIC value, however, it was very close to both the exponential and Weibull models.

So overall all the three parametric models are showing similar results.

But we think our COX PH model is the best one because all the assumptions are checked and the model is working good.