DST-CIMS, BHU

M.SC. STATISTICS & COMPUTING

SEM-IV PROJECT

SURVIVAL ANALYSIS: BREAST INVASIVE CARCINOMA

AKASH KUMAR GUPTA

JULY, 2021

ACKNOWLEDGEMENT

Place: Varanasi

Date: July 23, 2021

First, I wish to express my sincere gratitude to my supervisor, **Ass. Professor Mahavir Singh Panwar**, for his insightful comments, helpful information, practical advice, and unceasing ideas that have helped me at all times in my Project. His immense knowledge, profound experience, and professional expertise in Survival Analysis and Statistics have enabled me to complete this project successfully. Without his support and guidance, this project would not have been possible.

I would like to express my sincere gratitude to several individuals and **Banaras Hindu University** for supporting me throughout my Project.

I am ensuring that this project is finished by me.

Akash Kumar Gupta
M.Sc. Statistics & Computing
Sem-IV
DST-CIMS, BHU

CONTENTS

1. About Breast Cancer 5
2. About Survival Analysis 7
3. R Libraries Needed 8
4. Data Loading 8
4.1 Data Preprocessing 8
4.2 Variable Descriptions 9
5. Exploratory Data Analysis 10
5.1 Distribution of Vital Status 10
5.2 Distribution of type of Cancers only Prone to females 11
5.3 EDA for Breast Cancer 12
5.3.1 Distribution of Vital Status | Breast Cancer 12
5.3.2 Survival Plot concerning Age Category | Breast Cancer 13
5.3.3 Survival Plot concerning Race | Breast Cancer 14
5.3.4 Survival Plot concerning Ethnicity | Breast Cancer 15
5.3.5 Survival Plot concerning Therapy Type | Breast Cancer 16
5.3.6 Survival Plot concerning Cancer Stage | Breast Cancer 17
5.3.7 Survival Plot concerning Tumor Stage | Breast Cancer 18
5.3.8 Survival Plot concerning Lymph Node Stage | Breast Cancer 19
5.3.9 Survival Plot concerning Metastasis Stage | Breast Cancer 20
6. Analysis Phase 21
6.1 Comparison of different Cancers 21
Breast Cancer 24
6.2 Survival of Breast Cancer Patients without any covariates 24
6.3 Age 27
6.4 Race Category 31
6.5 Ethnicity Category 35
6.6 Therapy Type Category 39
6.7 Cancer Stage 44
6.8 Tumor Stage 48
6.9 Lymph Node Stage 52
6.10 Metastasis Stage 57
6.11 Multivariate Cox-Proportional Model 61
7. Results 63
8. Appendix: R Code 66
9. References 83

ABSTRACT

Breast Invasive Carcinoma, which is commonly known as Breast Cancer is most common in women. For American women, there is a 1 in 8 chance and For Indian women, there is a 1 in 29 chance that she will develop Breast Cancer in her lifetime.

In this project, we used Non-Parametric, Semi-Parametric, and Parametric methods to study the covariates associated with the Survival Time of Patients diagnosed with different stages of Breast Cancer.

Lastly, we have created a Semi-Parametric Model based on Univariate analysis results i.e., with covariates that significantly define the survival time of female Breast Cancer Patients.

We have found that covariate Age, Therapy Type, and Lymph Node Stage(N) can significantly contribute to predicting the Survival and Hazard of Breast Cancer Patients.

KEYWORDS

Survival Time

Censoring

Survival Function

Hazard Function

Probability Distributions

Kaplan Meier

Cox-Proportional Hazard

Life-Table

1. About Breast Cancer

What Is Breast Cancer?

Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control.

Breast cancer cells usually form a tumor that can often be seen on an x-ray or felt as a lump. Breast cancer occurs almost entirely in women, but men can get breast cancer, too.

It's important to understand that most breast lumps are benign and not cancer (malignant). Non-cancerous breast tumors are abnormal growths, but they do not spread outside of the breast. They are not life threatening, but some types of benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a health care professional to determine if it is benign or malignant (cancer) and if it might affect your future cancer risk.

Where breast cancer starts?

Breast cancers can start from different parts of the breast.

Most breast cancers begin in the ducts that carry milk to the nipple (ductal cancers) Some start in the glands that make breast milk (lobular cancers) There are also other types of breast cancer that are less common like phyllodes tumor and angiosarcoma A small number of cancers start in other tissues in the breast. These cancers are called sarcomas and lymphomas and are not really thought of as breast cancers. Although many types of breast cancer can cause a lump in the breast, not all do. See Breast Cancer Signs and Symptoms to learn what you should watch for and report to a health care provider. Many breast cancers are also found on screening mammograms, which can detect cancers at an earlier stage, often before they can be felt, and before symptoms develop.

Types of breast cancer?

There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumors and angiosarcoma are less common.

Once a biopsy is done, breast cancer cells are tested for proteins called estrogen receptors, progesterone receptors and HER2. The tumor cells are also closely looked at in the lab to find out what grade it is. The specific proteins found and the tumor grade can help decide treatment options.

How breast cancer spreads?

Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body.

The lymph system is a network of lymph (or lymphatic) vessels found throughout the body that connects lymph nodes (small bean-shaped collections of immune system cells). The clear fluid inside the lymph vessels, called lymph, contains tissue by-products and waste material, as well as immune system cells. The lymph vessels carry lymph fluid away from the breast. In the case of breast cancer, cancer cells can enter those lymph vessels and start to grow in lymph nodes. Most of the lymph vessels of the breast drain into:

Lymph nodes under the arm (axillary nodes) Lymph nodes around the collar bone (supraclavicular [above the collar bone] and infraclavicular [below the collar bone] lymph nodes) Lymph nodes inside the chest near the breast bone (internal mammary lymph nodes)

If cancer cells have spread to your lymph nodes, there is a higher chance that the cells could have traveled through the lymph system and spread (metastasized) to other parts of your body. The more lymph nodes with breast cancer cells, the more likely it is that the cancer may be found in other organs. Because of this, finding cancer in one or more lymph nodes often affects your treatment plan. Usually, you will need surgery to remove one or more lymph nodes to know whether the cancer has spread.

Still, not all women with cancer cells in their lymph nodes develop metastases, and some women with no cancer cells in their lymph nodes develop metastases later.

You also may see or hear certain words used to describe the stage of the breast cancer:

Local: The cancer is confined within the breast.

Regional: The lymph nodes, primarily those in the armpit, are involved.

Distant: The cancer is found in other parts of the body as well.

Information about the TNM staging system

1. The **T** (size) category describes the original (primary) tumor:

   TX means the tumor can't be assessed.

   T0 means there isn't any evidence of the primary tumor.

   T is means the cancer is "in situ" (the tumor has not started growing into healthy breast tissue).

   T1, T2, T3, T4: These numbers are based on the size of the tumor and the extent to which it has grown into neighboring breast tissue. The higher the T number, the larger the tumor and/or the more it may have grown into the breast tissue.

2. The **N** (lymph node involvement) category describes whether or not the cancer has reached nearby lymph nodes:

   NX means the nearby lymph nodes can't be assessed, for example, if they were previously removed.

   N0 means nearby lymph nodes do not contain cancer.

   N1, N2, N3: These numbers are based on the number of lymph nodes involved and how much cancer is found in them. The higher the N number, the greater the extent of the lymph node involvement.

3. The **M** (metastasis) category tells whether or not there is evidence that the cancer has traveled to other parts of the body:

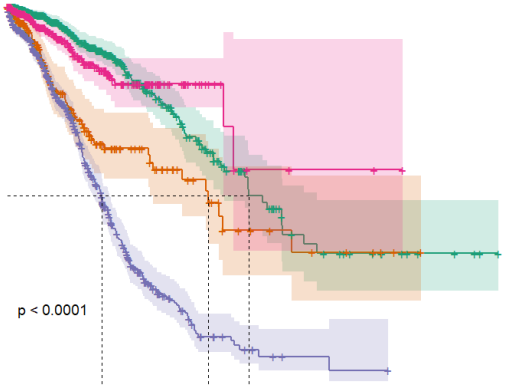
   MX means metastasis can't be assessed.

   M0 means there is no distant metastasis.

   M1 means that distant metastasis is present.

2. About Survival Analysis

Survival analysis corresponds to a set of statistical approaches used to investigate the time it takes for an event of interest to occur.



Survival time and type of events in cancer studies.

There are different types of events, including:

- Relapse
- Progression
- Death

The two most important measures in cancer studies include:

I. the time to death; and

II. the relapse-free survival time, which corresponds to the time between response to treatment and recurrence of the disease. It's also known as disease-free survival time and event-free survival time.

Censoring

As mentioned above, survival analysis focuses on the expected duration of time until occurrence of an event of interest (relapse or death). However, the event may not be observed for some individuals within the study time period, producing the so-called censored observations.

Censoring may arise in the following ways:

I. a patient has not (yet) experienced the event of interest, such as relapse or death, within the study time period;

II. a patient is lost to follow-up during the study period;

III. a patient experiences a different event that makes further follow-up impossible.

This type of censoring, named right censoring, is handled in survival analysis.

Survival and hazard functions

Two related probabilities are used to describe survival data: the *survival probability* and the *hazard probability*.

The survival probability, also known as the survivor function ***S(t)***, is the probability that an individual survives from the time origin (e.g., diagnosis of cancer) to a specified future time t.

The hazard, denoted by ***h(t)***, is the probability that an individual who is under observation at a time t has an event at that time.

Note that, in contrast to the survivor function, which focuses on not having an event, the hazard function focuses on the event occurring.

3. R Libraries Needed

```
# if (!requireNamespace("BiocManager", quietly = TRUE))
#   install.packages("BiocManager")
# BiocManager::install()
library(RTCGA)
library(RTCGA.clinical)
library(RTCGA.mRNA)
library(survminer)
library(survival)
library(SurvRegCensCov)
library(flexsurv)
library(tidyverse)
library(pivottabler)
library(writexl)
library(ggplot2)
library(Hmisc)
```

4. Data Loading

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

With the help of the BiocManager and other libraries of R, we have extracted few Required variables from UCEC.clinical, BRCA.clinical, OV.clinical, CESC.clinical for our analysis.

Contains Survival Information of 2528 Patients having –

BRCA - Breast invasive carcinoma
CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma
OV - Ovarian serous cystadenocarcinoma
UCEC - Uterine Corpus Endometrial Carcinoma

times	bcr_patient_barcode	patient.vital_status	admin.disease_code	patient.drugs.drug.therapy_types.therapy_type
1065	TCGA-2E-A9G8	0	ucec	chemotherapy
0	TCGA-4E-A92E	0	ucec	
883	TCGA-5B-A90C	0	ucec	chemotherapy
33	TCGA-5S-A9Q8	0	ucec	chemotherapy
3251	TCGA-A5-A0G1	1	ucec	
4054	TCGA-A5-A0G2	0	ucec	
1079	TCGA-A5-A0G3	0	ucec	chemotherapy
790	TCGA-A5-A0G5	0	ucec	

Contains Demographic, Survival, and Pathologic Information of 1098 BRCA - Breast Cancer Patients -

times	bcr_patient_barcode	atient.vital_stat	patient.gender	patient.race	patient.ethnicity	ent.days_to_b	drug.therapy_types	_event.path	ories.path	gories.path	ories.patho
3767	TCGA-3C-AAAU	0	female	white	not hispanic or latino	-20211	chemotherapy	stage x	tx	nx	mx
3801	TCGA-3C-AALI	0	female	black or african american	not hispanic or latino	-18538		stage iib	t2	n1a	m0
1228	TCGA-3C-AALJ	0	female	black or african american	not hispanic or latino	-22848		stage iib	t2	n1a	m0
1217	TCGA-3C-AALK	0	female	black or african american	not hispanic or latino	-19074	chemotherapy	stage ia	t1c	n0 (i+)	m0
158	TCGA-4H-AAAK	0	female	white	not hispanic or latino	-18371	chemotherapy	stage iiia	t2	n2a	m0
1477	TCGA-5L-AAT0	0	female	white	hispanic or latino	-15393	hormone therapy	stage iia	t2	n0	m0
1471	TCGA-5L-AAT1	0	female	white	hispanic or latino	-23225	hormone therapy	stage iv	t2	n0	m1
12	TCGA-5T-A9QA	0	female	black or african american	not hispanic or latino	-19031		stage iia	t2	nx	mx

4.1 Data Preprocessing

I. Shortened the long variable names.

II. Converted Survival Time from the number of days to the number of Years.

III. Converted Patient.days_to_birth variable to Age Variable.

IV. Categorized the age Variable into 3 Categories –

Young (0-40 Years), Middle (40-60 Years), and Old (60+ Years).

V. We have the right-censored data, so had to remove few observations having negative survival time.

VI. Removed 12 Male Breast Cancer Patients' data.

VII. Removed Race: american indian or alaska native as it has only 1 observation.

VIII. Grouped subgroups of Therapy Type in 4 groups-

Chemotherapy, Hormone therapy, No Information, and Other

IX. Grouped subgroups of Cancer Stage in 5 Stages- Stage 1,2,3,4, and X.

X. Grouped subgroups of Tumor Stage in 5 Stages- Stage 1,2,3,4, and X.

XI. Grouped subgroups of Lymph Node Stage in 5 Stages- Stage 0,1,2,3, and X.

XII. Grouped subgroups of Metastasis Stage in 3 Stages- Stage 0,1, and X.

4.2 Variable Descriptions

Survival Time	How long Patient Survives after first diagnosis. (Years)	Numeric
Patient_Code	Code which uniquely identifies each Cancer Patient.	Factor
Vital_Status	Patient is alive or not; 0 - Alive; 1 - Not Alive (Event)	Factor
Gender	Patient's Sex	Factor
Race	Patient's Race	Factor
Ethnicity	Patient's Ethnicity	Factor
Age	Patients's Age in Number of Years	Numeric
Therapy_Type	Breast Cancer Therapy taken by Patient	Factor
Cancer_Stage	Patient's Breast Cancer Stage- 1,2,3,4, and X	Factor
Tumor_Stage	Patient's Tumor Stage - 1,2,3,4, and X	Factor
Lymph_Node_Stage	Pateint's Lymph Node Stage - 0,1,2,3, and X	Factor
Metastasis_Stage	Patient's Metastasis Stage- 0,1, and X	Factor
Age_Category	Patients Age Category – Young (0-40); Middle (40-60); Old (60+)	Factor
Disease_Code	Whether Patient have BRCA or CESC or OV or UCEC	Factor

Final Dataset for Survival Comparison of Breast Cancer with UCEC, OV, CESC - “clin_can_data”

Survival_Time	Patient_code	Vital_Status	Disease_Code	Therapy_Type
2.92	TCGA-2E-A9G8	0	ucec	chemotherapy
2.42	TCGA-5B-A90C	0	ucec	chemotherapy
0.09	TCGA-5S-A9Q8	0	ucec	chemotherapy
8.91	TCGA-A5-A0G1	1	ucec	No Information
11.11	TCGA-A5-A0G2	0	ucec	No Information
2.96	TCGA-A5-A0G3	0	ucec	chemotherapy
2.16	TCGA-A5-A0G5	0	ucec	No Information
5.93	TCGA-A5-A0G9	0	ucec	No Information

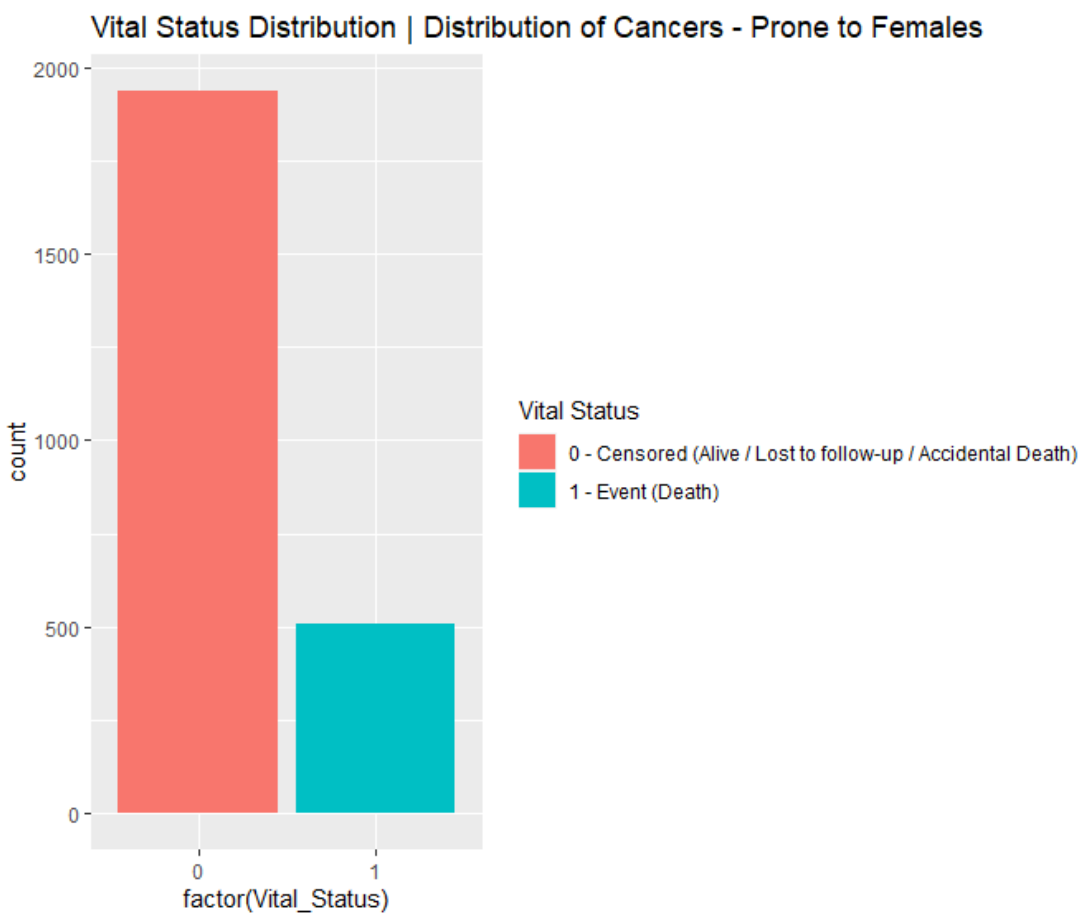
Final Dataset for Survival Analysis of Breast Cancer Patients - “BRCA_data”

Survival_Time	Patient_code	Vital_Status	Gender	Race	Ethnicity	Age	Therapy_Type	Cancer_Stage	Tumor_Stage	Lymph__node_Stage	Metastasis_Stage	Age_Category
10.32	TCGA-3C-AAAU	0	female	white	not hispanic or latino	55.37	chemotherapy	Stage x	Tumor Stage x	Lymph Node Stage x	Metastatis Stage x	Middle Age
10.41	TCGA-3C-AALI	0	female	black or african american	not hispanic or latino	50.79	No Information	Stage 2	Tumor Stage 2	Lymph Node Stage 1	Metastatis Stage 0	Middle Age
3.36	TCGA-3C-AALJ	0	female	black or african american	not hispanic or latino	62.6	No Information	Stage 2	Tumor Stage 2	Lymph Node Stage 1	Metastatis Stage 0	Old Age
3.33	TCGA-3C-AALK	0	female	black or african american	not hispanic or latino	52.26	chemotherapy	Stage 1	Tumor Stage 1	Lymph Node Stage 0	Metastatis Stage 0	Middle Age
0.43	TCGA-4H-AAAK	0	female	white	not hispanic or latino	50.33	chemotherapy	Stage 3	Tumor Stage 2	Lymph Node Stage 2	Metastatis Stage 0	Middle Age
4.05	TCGA-5L-AAT0	0	female	white	hispanic or latino	42.17	hormone therapy	Stage 2	Tumor Stage 2	Lymph Node Stage 0	Metastatis Stage 0	Middle Age
4.03	TCGA-5L-AAT1	0	female	white	hispanic or latino	63.63	hormone therapy	Stage 4	Tumor Stage 2	Lymph Node Stage 0	Metastatis Stage 1	Old Age
0.03	TCGA-5T-A9QA	0	female	black or african american	not hispanic or latino	52.14	No Information	Stage 2	Tumor Stage 2	Lymph Node Stage x	Metastatis Stage x	Middle Age

5. Exploratory Data Analysis

- We will firstly visualize Breast Cancer Patients Data with Other 3 Cancers.
- Then we will focus on the Survival Data for Breast Cancer Patients.

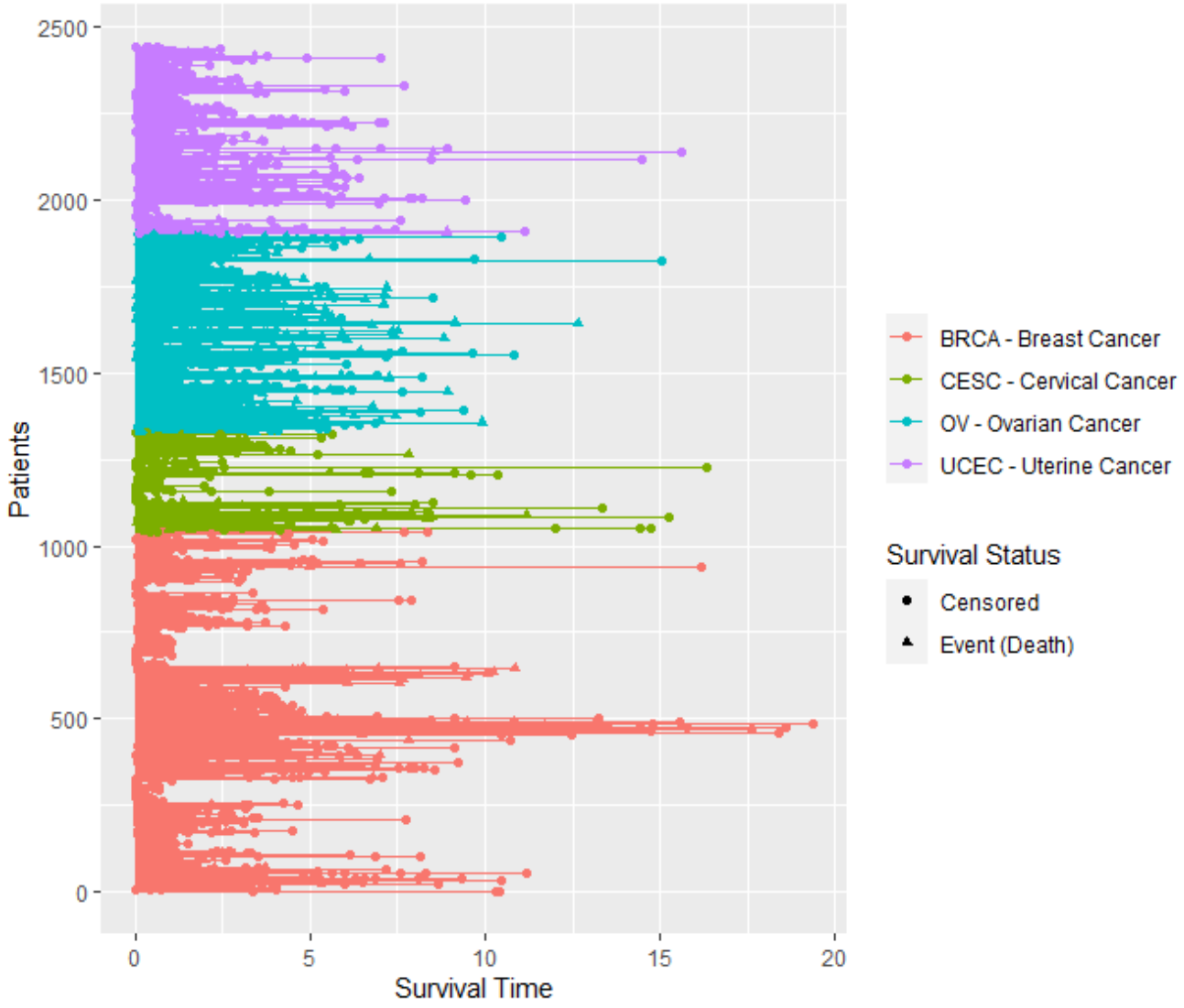
5.1 Distribution of Vital Status



In our data we only have around 25% confirmed Survival time (Death) and rest of the survival time is censored.

5.2 Distribution of type of Cancers only Prone to females

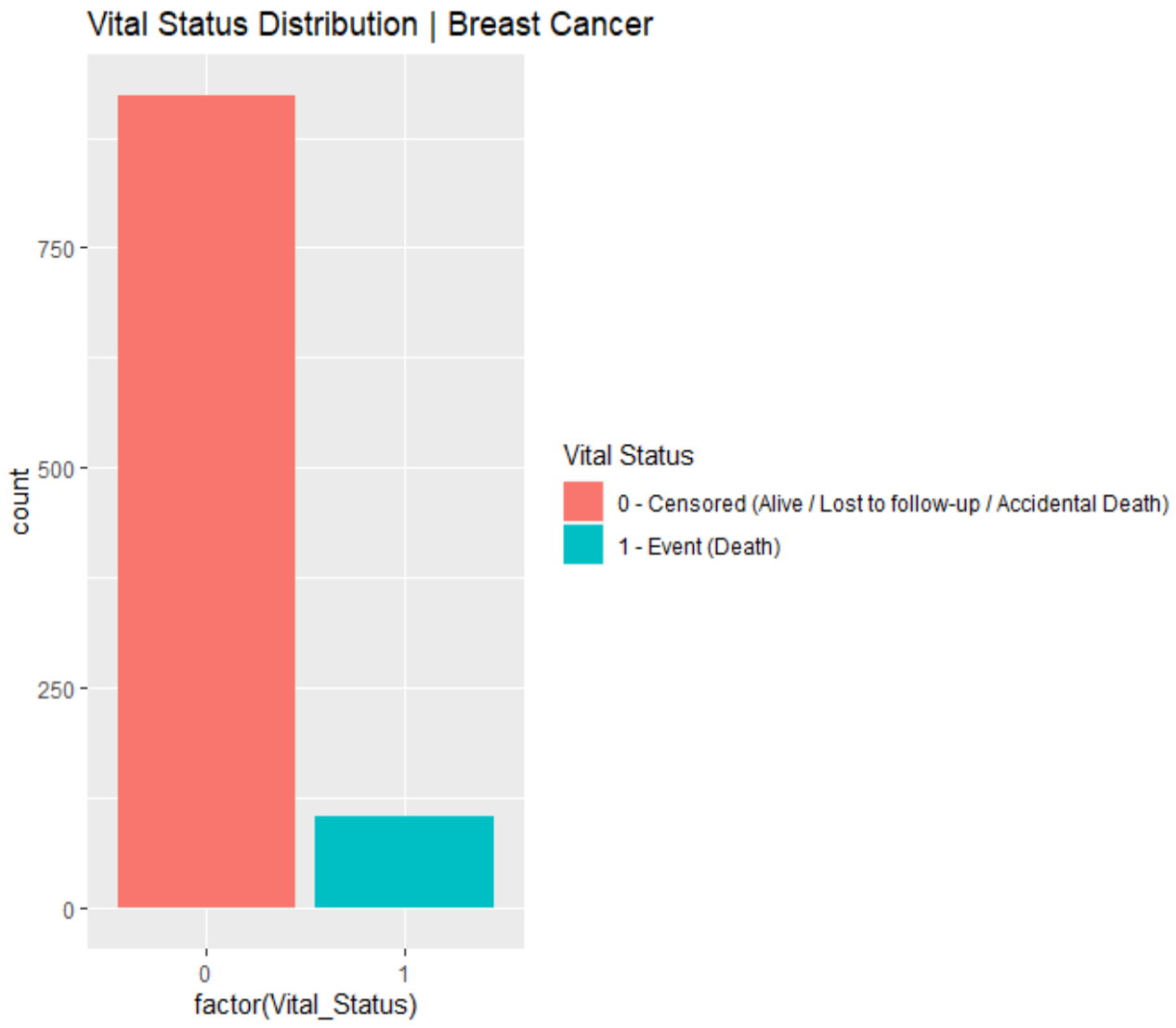
```
##
##     brca     cesc       ov     ucec
## 42.57061 11.74785 23.53664 22.14490

##
```



We can see out these four cancer types, 42% cases were alone having Breast Cancer and Cervical Cancer is the least prone out of these.

5.3 EDA for Breast Cancer

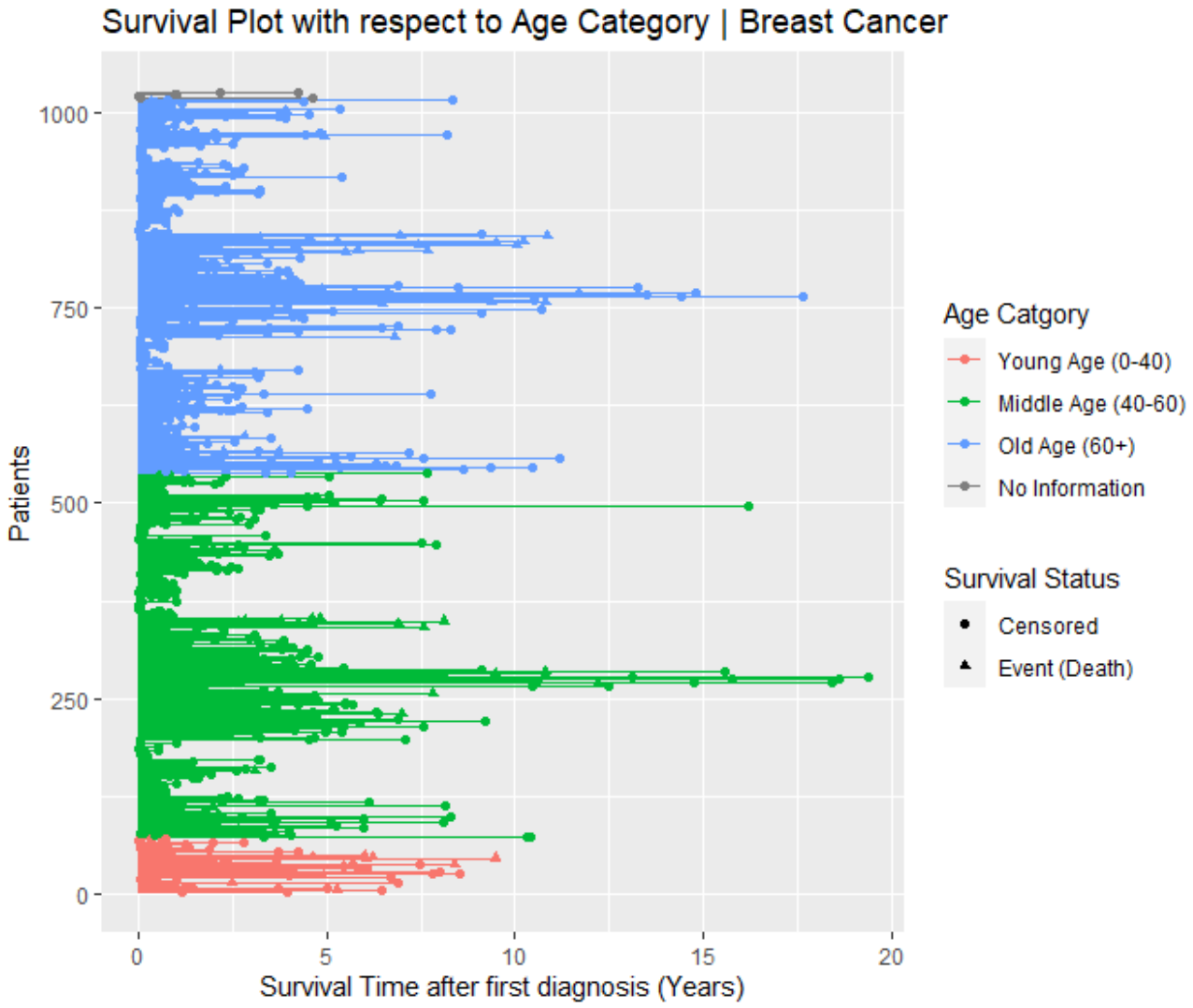
5.3.1 Distribution of Vital Status | Breast Cancer



In our data we only have around 10% confirmed Survival time (Death) and rest 90% of the survival time is censored.

5.3.2 Survival Plot concerning Age Category | Breast Cancer

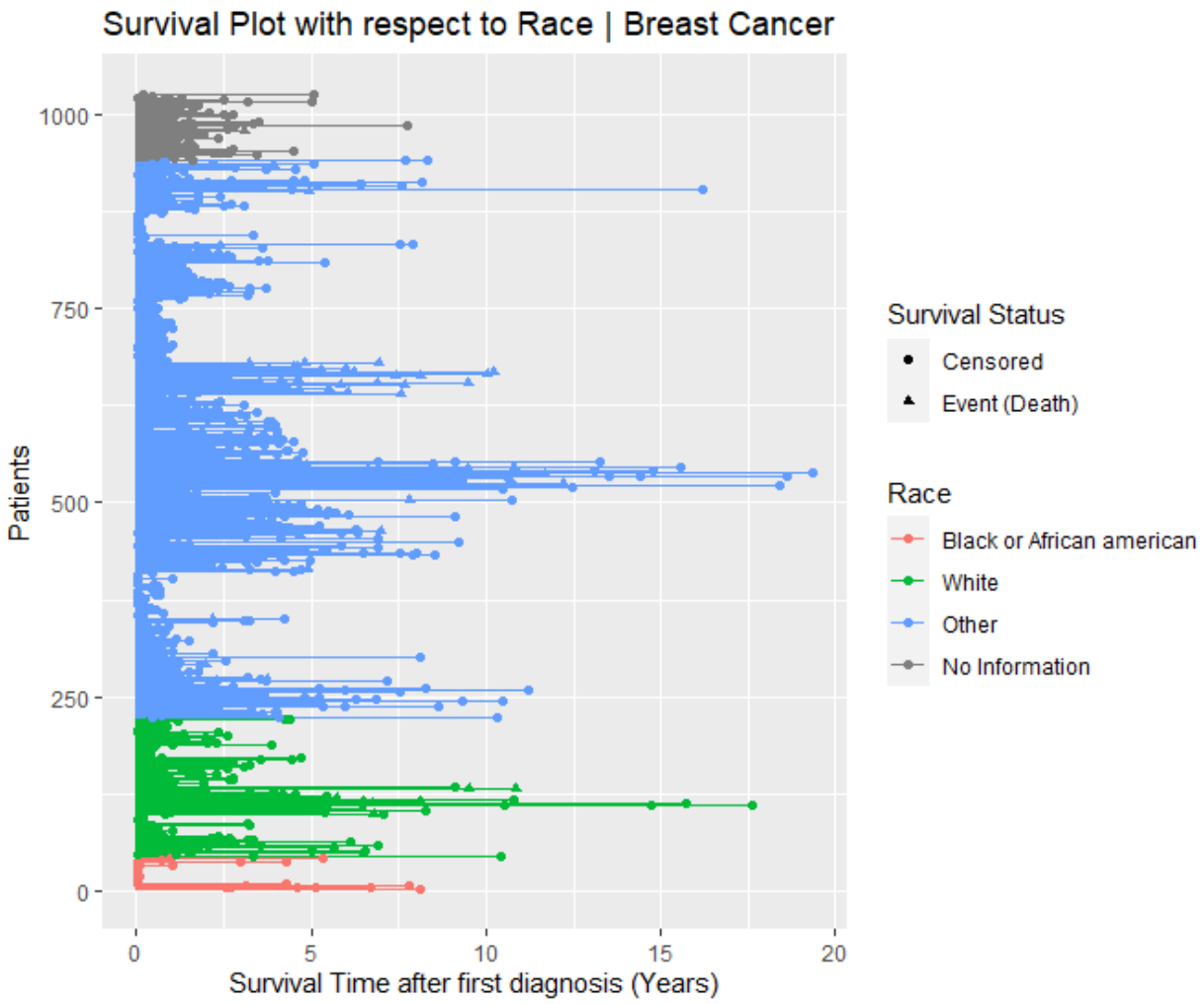
	Young Age	Middle Age	Old Age	NA	Total
0	59	433	422	9	923
1	11	35	58		104
Total	70	468	480	9	1027



We Can see that Young People (0-40) are Very Less affected by Breast Cancer in comparison to 40+ Age Groups This implies that, Older females are more Prone to Breast Cancer.

5.3.3 Survival Plot concerning Race | Breast Cancer

	asian	black or african american	white	NA	Total
0	41	160	642	80	923
1	1	19	79	5	104
Total	42	179	721	85	1027

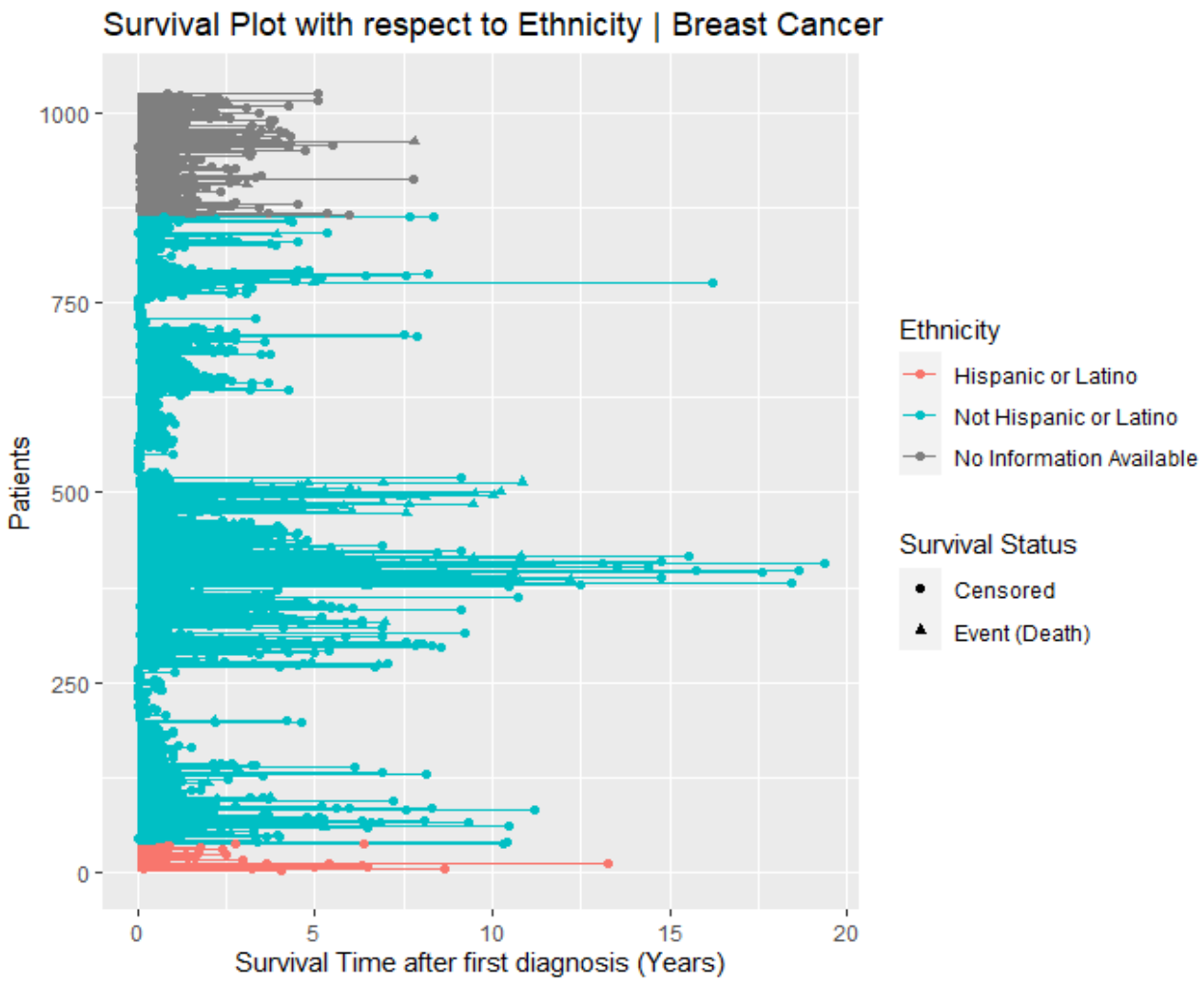


As this data is of US citizens and we know that White people constitute 77% of the population and Black or African American constitutes 13% of the population of US.

But in our data white people constitute 70% only and Black or African people constitute 17% of Cancer patients. Hence, we can say that Black females are more prone to Breast Cancer in comparison to white females.

5.3.4 Survival Plot concerning Ethnicity | Breast Cancer

	hispanic or latino	not hispanic or latino	NA	Total
0	37	732	154	923
1		96	8	104
Total	37	828	162	1027

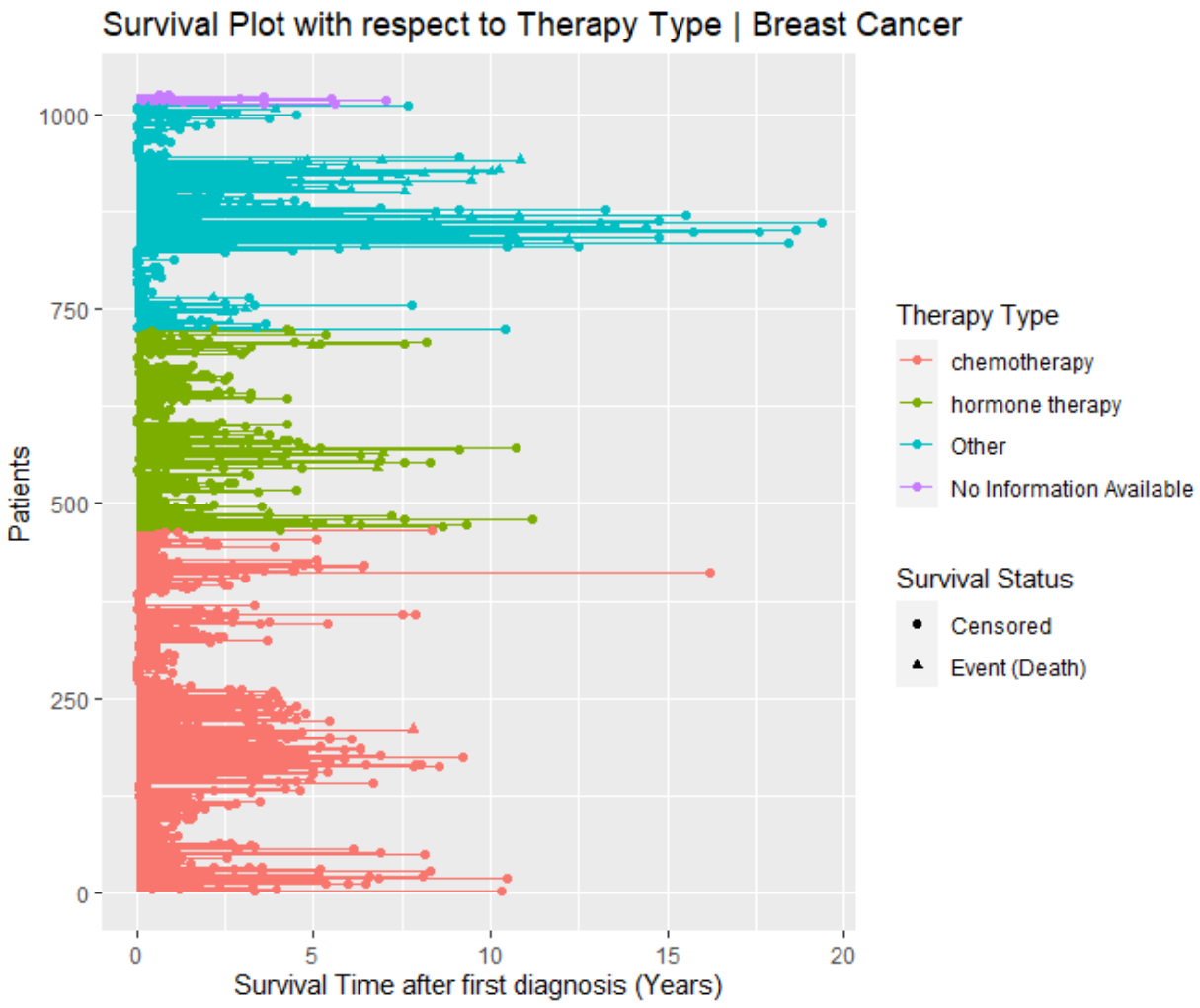


For 15% females we don't have their respective ethnicity information. We don't have any Non censored (Event-Death) observation for Hispanic or Latino i.e. No Hispanic or Latino breast cancer patient died within the observation period. In US Non-Hispanic or Latino population constitutes 60% of the population but in our data this Ethnicity constitutes 80% of the total female breast cancer patients.

While Hispanic or Latino Population constitutes 18% of the population in US but in our data only 3% patients have this ethnicity. Hence, we can say that, comparatively, Hispanic and Latino females are less prone to Breast Cancer.

5.3.5 Survival Plot concerning Therapy Type | Breast Cancer

	chemotherapy	hormone therapy	No Information	Other	Total
0	447	253	210	13	923
1	18	7	78	1	104
Total	465	260	288	14	1027

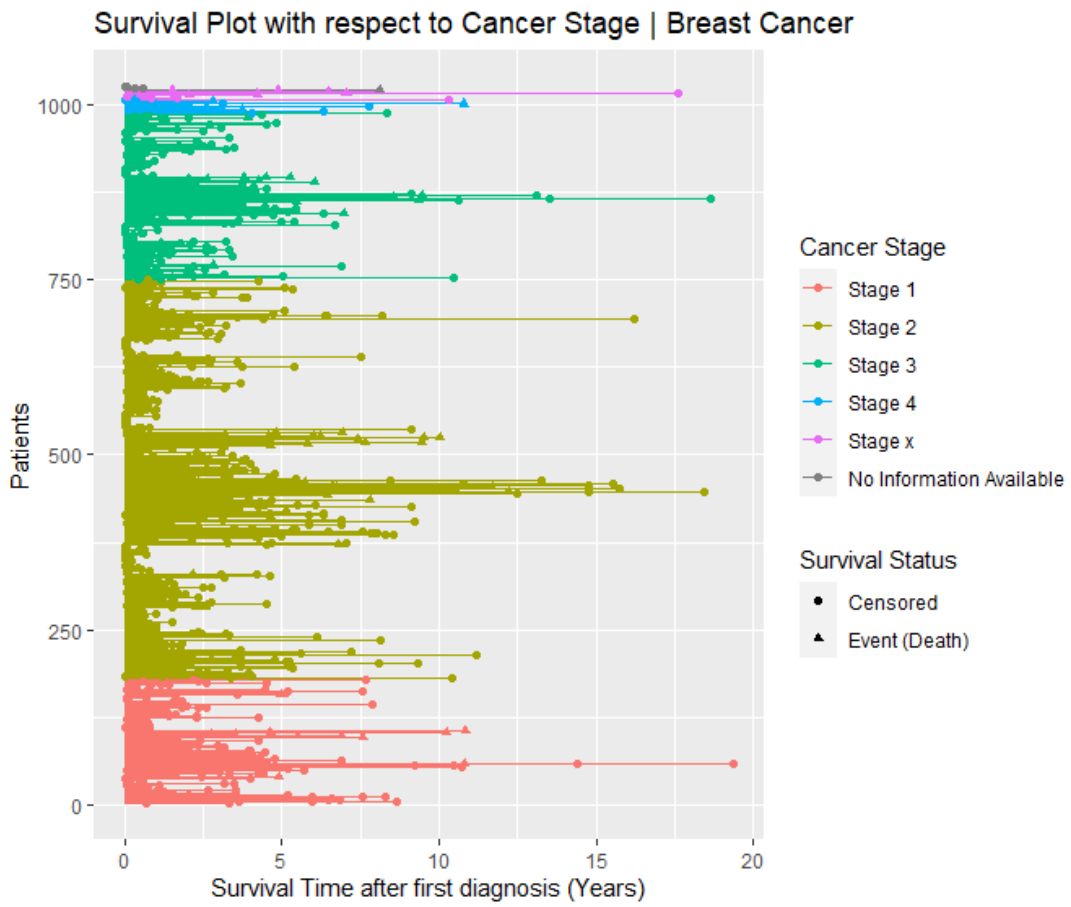


For 28% females we don't have their therapy type information.

Comparatively, a greater number of females with Breast Cancer are having chemotherapy followed by Hormone therapy.

5.3.6 Survival Plot concerning Cancer Stage | Breast Cancer

	Stage 1	Stage 2	Stage 3	Stage 4	Stage x	NA	Total
0	166	527	209	10	7	4	923
1	13	44	30	9	7	1	104
Total	179	571	239	19	14	5	1027



Most of the females are diagnosed with 'Stage 2' Breast Cancer (55%), followed by 'Stage 3' (23%) and 'Stage 1' (17%).

All patients diagnosed with 'Stage x' are non-censored (Death within clinical trial interval)

90% patients diagnosed with 'Stage 4' breast cancer are non-censored (Death within clinical trial interval)

12% patients diagnosed with 'Stage 3' breast cancer are non-censored (Death within clinical trial interval)

Therapy Type ~ Cancer Stage

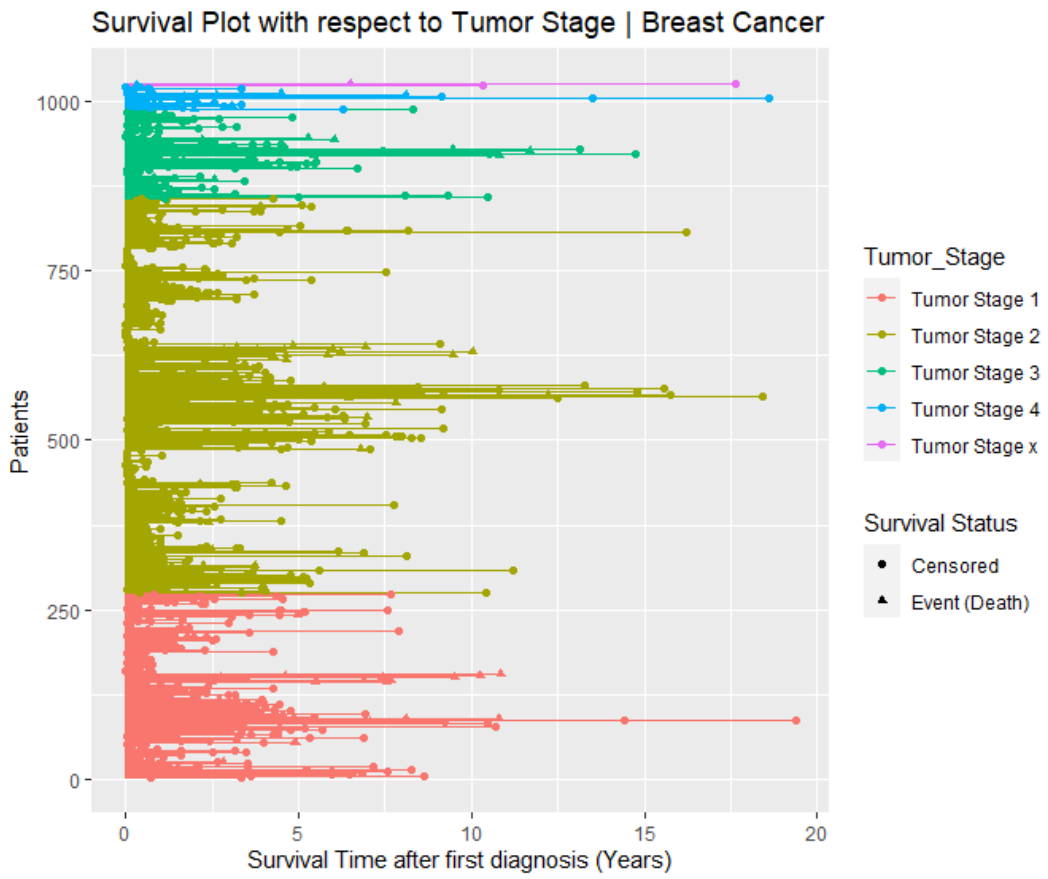
	Stage 1	Stage 2	Stage 3	Stage 4	Stage x	NA	Total
chemotherapy	56	264	135	4	2	4	465
hormone therapy	62	152	38	5	3		260
No Information	60	146	63	9	9	1	288
Other	1	9	3	1			14
Total	179	571	239	19	14	5	1027

For 'Stage 1' Breast Cancer Both 'Chemotherapy' and 'Hormone Therapy' are used but Still 'Hormone Therapy' is preferred slightly more.

In 'Stage 2' and 'Stage 3', 'Chemotherapy' is Preferred.

5.3.7 Survival Plot concerning Tumor Stage | Breast Cancer

	Tumor Stage 1	Tumor Stage 2	Tumor Stage 3	Tumor Stage 4	Tumor Stage x	Total
0	247	533	115	26	2	923
1	26	51	16	10	1	104
Total	273	584	131	36	3	1027



56% of females are diagnosed with breast cancer while having 'Stage 2 Tumor'

26% of females are diagnosed with breast cancer while having 'Stage 1 Tumor'

Only 3% of females are diagnosed with breast cancer while having 'Stage 4 Tumor'

27% of females having 'Stage 4 Tumor' died within clinical trial time interval (Not censored)
i.e. Comparatively, females with 'Tumor Stage 4' have high death risk.

Cancer Stage ~ Tumor Stage

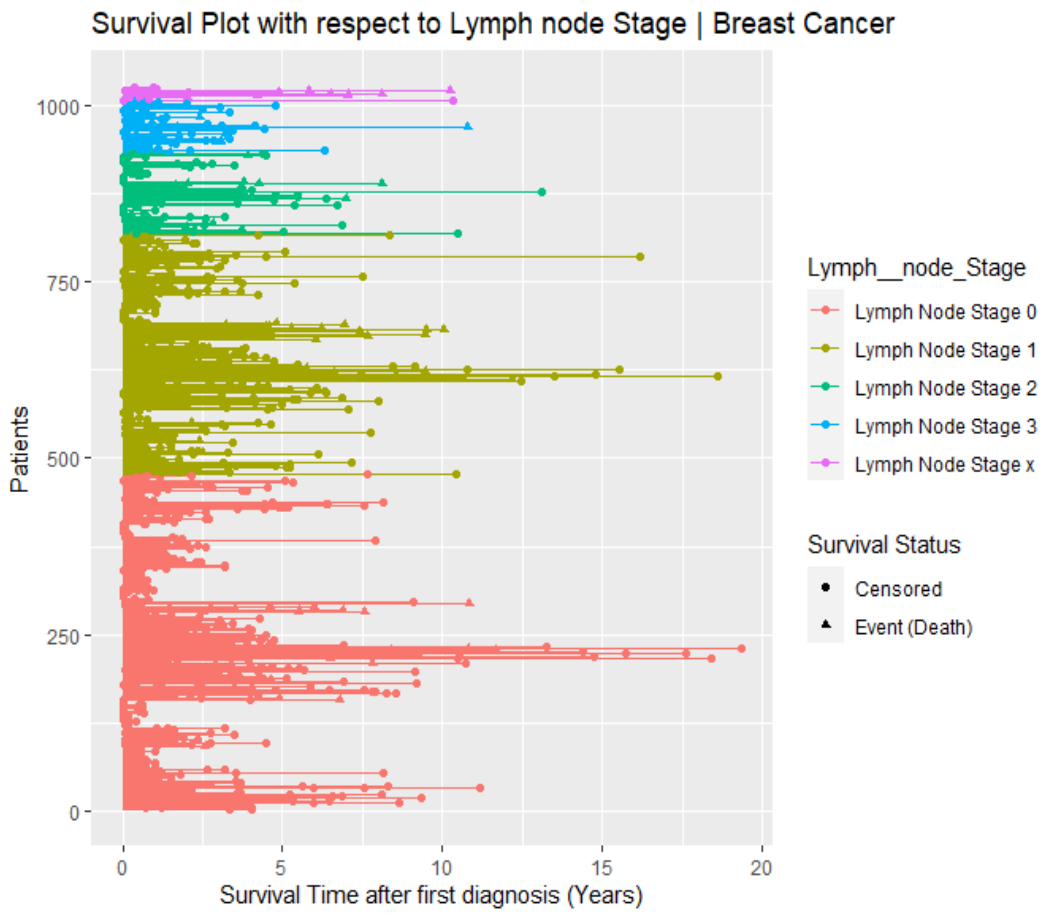
	Tumor Stage 1	Tumor Stage 2	Tumor Stage 3	Tumor Stage 4	Tumor Stage x	Total
Stage 1	179					179
Stage 2	64	468	39			571
Stage 3	23	102	88	26		239
Stage 4	1	8	4	6		19
Stage x	5	4		2	3	14
NA	1	2		2		5
Total	273	584	131	36	3	1027

In 'Stage 1' Breast Cancer, only 'Tumor Stage 1' is possible.
In 'Stage 2' Breast Cancer, 'Tumor Stage 1,2,3' are possible but mostly 'Tumor Stage 2' is found. In 'Stage 3' Breast Cancer, all 4 Tumor Stages are possible but mostly 'Tumor Stage 2,3' are found.

In 'Stage 4' Breast Cancer, 'Tumor Stage 1' are is less likely.

5.3.8 Survival Plot concerning Lymph Node Stage | Breast Cancer

	Lymph Node Stage 0	Lymph Node Stage 1	Lymph Node Stage 2	Lymph Node Stage 3	Lymph Node Stage x	Total
0	449	299	101	63	11	923
1	28	42	15	10	9	104
Total	477	341	116	73	20	1027



In 46% Cases Lymph Node are not involved.

When Nearby Lymph Nodes cannot be assessed, 45% females died within Clinical trial period.

Cancer Stage ~ Lymph Node Stage

	Lymph Node Stage 0	Lymph Node Stage 1	Lymph Node Stage 2	Lymph Node Stage 3	Lymph Node Stage x	Total
Stage 1	170	6			3	179
Stage 2	297	270			4	571
Stage 3	5	58	110	65	1	239
Stage 4	1	5	4	6	3	19
Stage x	3	1		2	8	14
NA	1	1	2		1	5
Total	477	341	116	73	20	1027

For ‘Stage 1’ Cancer, Only ‘Lymph Node Stage 0’ or ‘Lymph Node Stage 1’ are possible, but mostly ‘Lymph Node Stage 0’ is found.

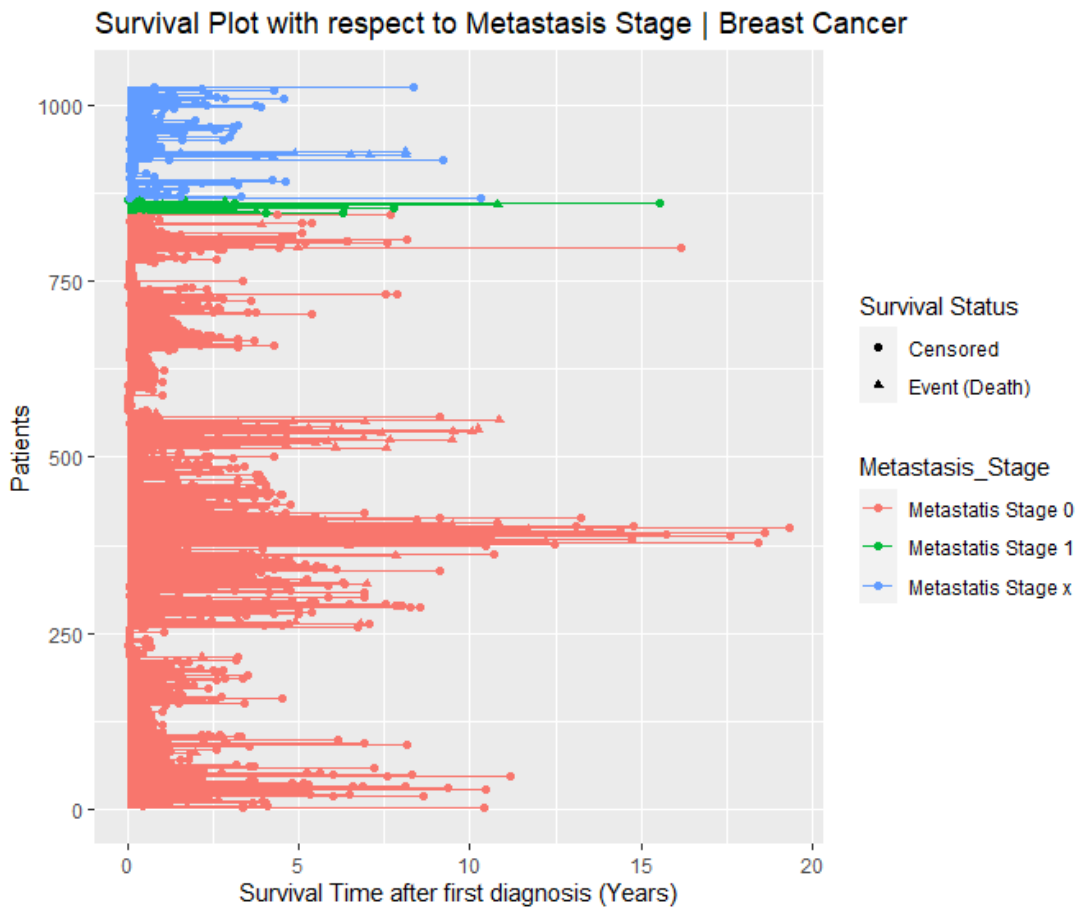
For ‘Stage 2’ Cancer, Only ‘Lymph Node Stage 1’ or ‘Lymph Node Stage 2’ are possible, almost equally.

For ‘Stage 3’ Cancer, All ‘Lymph Node Stages 0,1,2,3 are possible, but mostly ’Lymph Node Stages 2,3’.

For ‘Stage 4’ Cancer, All ‘Lymph Node Stages 0,1,2,3 are possible, but mostly ’Lymph Node Stages 3’.

5.3.9 Survival Plot concerning Metastasis Stage | Breast Cancer

	Metastatis Stage 0	Metastatis Stage 1	Metastatis Stage x	Total
0	759	12	152	923
1	87	9	8	104
Total	846	21	160	1027



In 82% of the females diagnosed with Breast Cancer no distant spread is found i.e. 'Metastasis Stage 0'.

In rest 15% cases, Distant spread (metastasis) cannot be assessed i.e. 'Metastasis Stage 0'

Only in less than 3% Cases, 'Metastasis Stage 1' is found.

Cancer Stage ~ Metastasis Stage

	Metastatis Stage 0	Metastatis Stage 1	Metastatis Stage x	Total
Stage 1	160		19	179
Stage 2	486	1	84	571
Stage 3	193	1	45	239
Stage 4		19		19
Stage x	3		11	14
NA	4		1	5
Total	846	21	160	1027

We can easily observe that in 'Stage 4' Cancer we almost every time find 'Metastasis Stage 1' cancer.

i.e. In Cancer Stages 1,2,3 we only find 'Metastasis Stage 0 or X'.

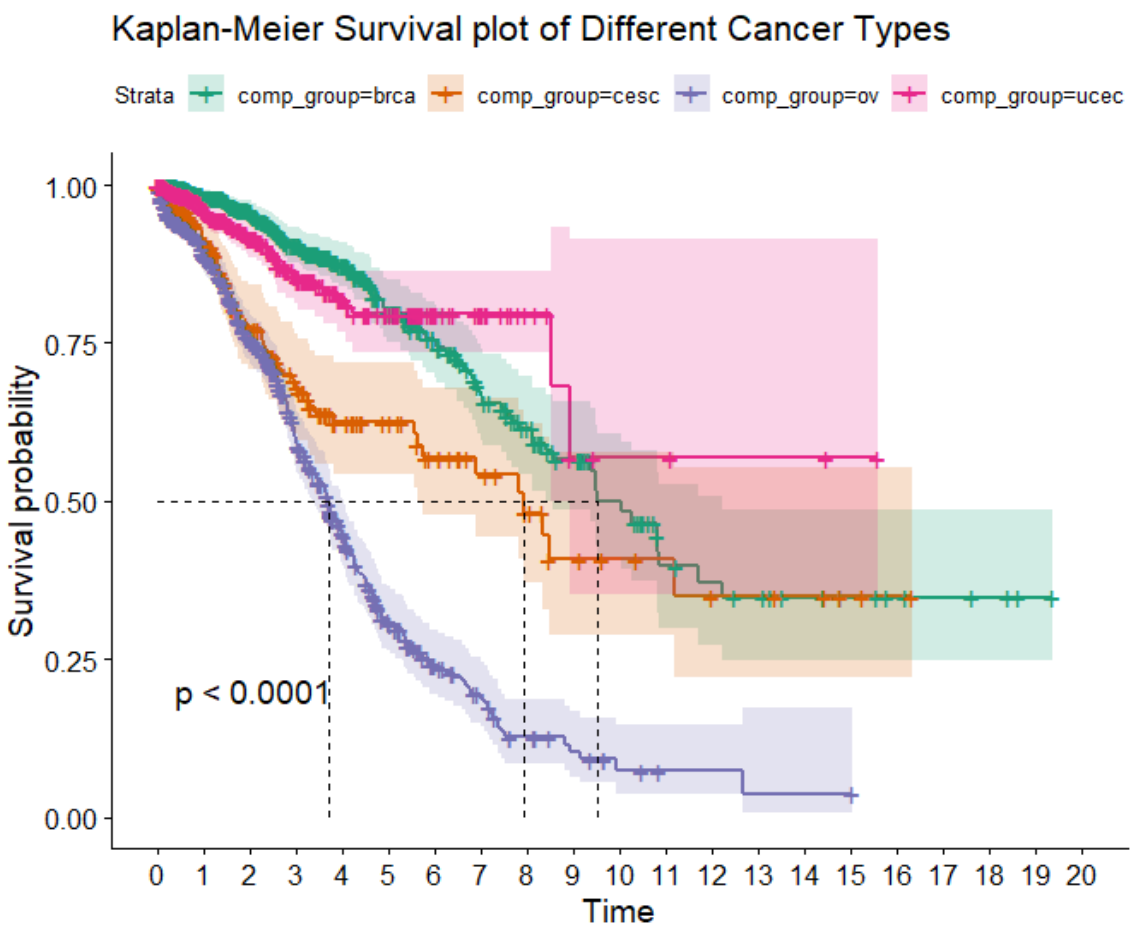
6. Analysis Phase

- Firstly, we will fit univariate Non-Parametric, Semi-Parametric, and Parametric Survival models to find how these covariates affects the Survival and Hazard.
- Then we will try to fit multivariate Survival model based on Univariate analysis results.

6.1 Comparison of different Cancers

```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                      n events median 0.95LCL 0.95UCL
## comp_group=brca 1040    104   9.51    8.56   12.21
## comp_group=cesc  287     60   7.91    5.62      NA
## comp_group=ov    575    297   3.71    3.37    4.03
## comp_group=ucec  541     45     NA    8.53      NA
```

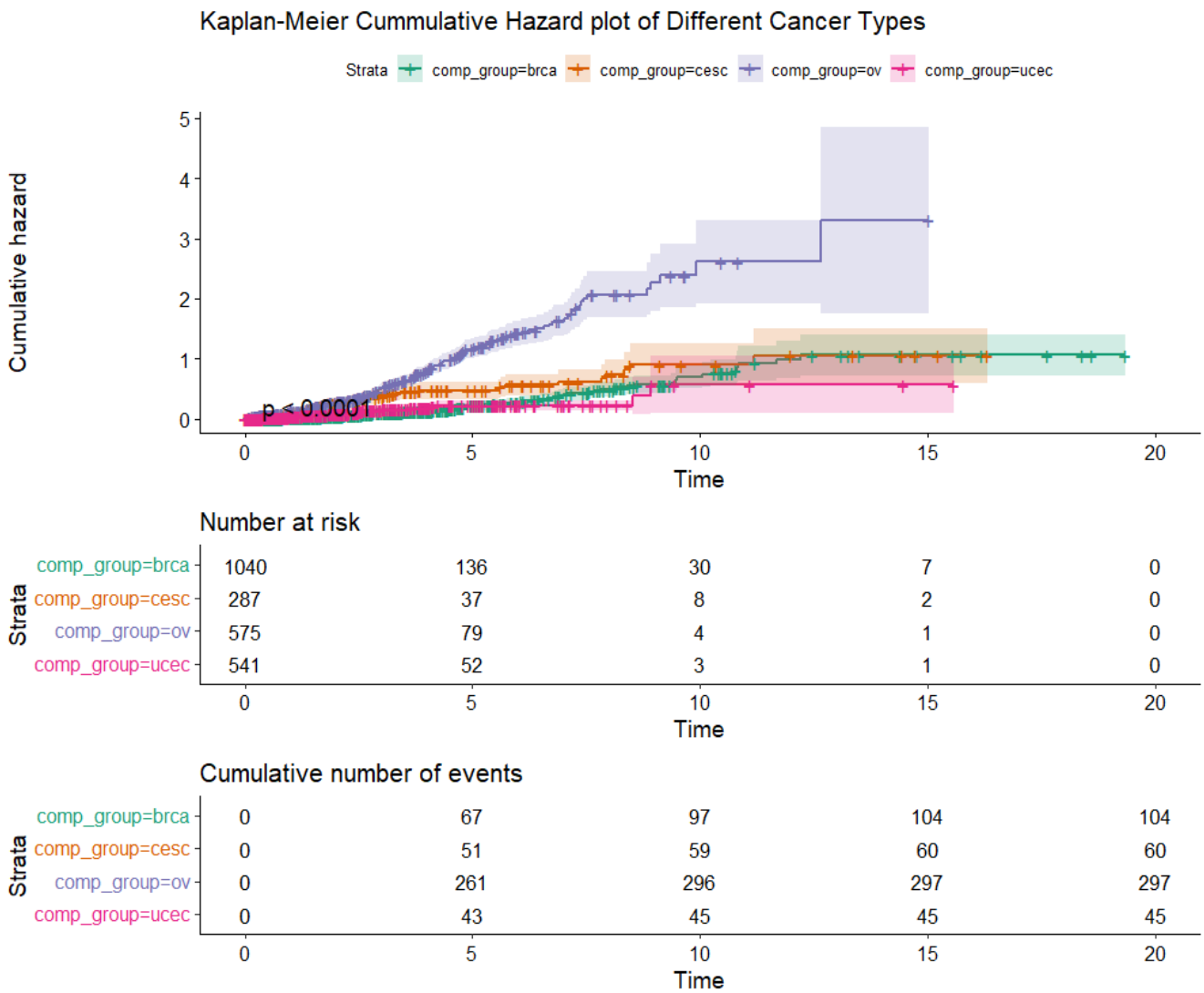
6.1.1



Survival Curves seems to be significantly different from each other. We are not able to estimate the Median survival time for UCEC using the K-M model as the probability of survival didn't reach 0.50.

Median survival time for BRCA, 9.51 Years which is larger than that of CESC, 7.91 Years and OV having the least, 3.71 Years.

6.1.2



We can observe, Cumulative Hazard increases much quickly for OV as compared to BRCA, CESC, and UCEC.

6.1.3 Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Disease_Code, data = cancer_clin_data)
##
##                      N Observed Expected (O-E)^2/E (O-E)^2/V
## Disease_Code=brca 1040      104    221.8    62.580   112.933
## Disease_Code=cesc  287       60     54.9     0.472     0.533
## Disease_Code=ov    575      297    138.8   180.403   251.958
## Disease_Code=ucec  541       45     90.5    22.875    28.119
##
##  Chisq= 269  on 3 degrees of freedom, p= <2e-16
```

To have the same survival curves, BRCA and UCEC are expected to have more number of events when CESC and OC are expected to have a lesser number of events. P-value is less than 0.05 at a 95% confidence level, which means there is a significant difference in the survival curves.

6.1.4 Pairwise-Difference in Survival Curves

```
##      brca cesc ov
## cesc ****
## ov   **** ****
## ucec      **** ****
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

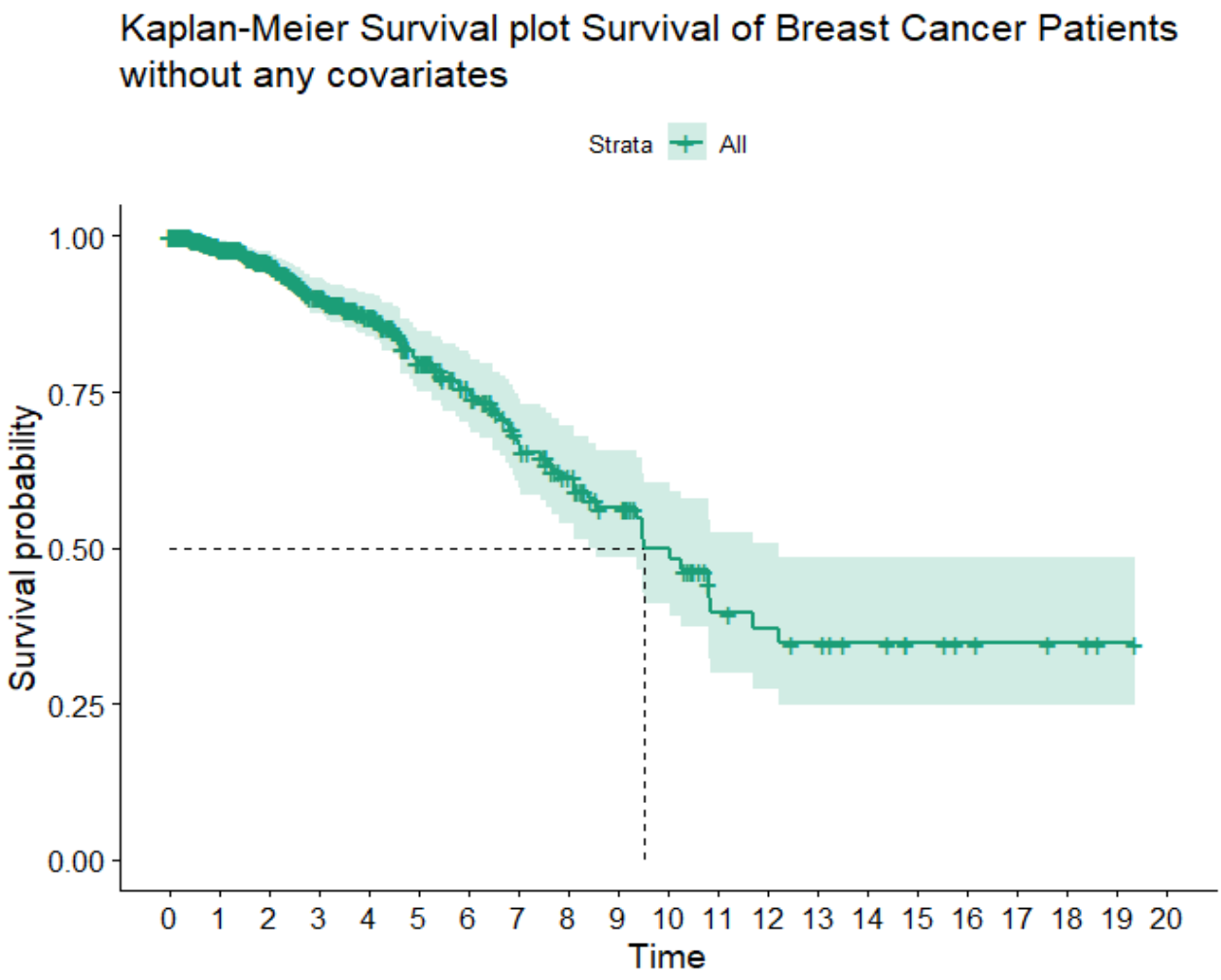
We can’t find any significant difference in survival curves of CESC and OV. Same goes with the UCEC and BRCA. Rest other pairs have significant difference in their survival curves.

Breast Cancer

6.2 Survival of Breast Cancer Patients without any covariates

6.2.1 Non-Parametric Null Model

```
##
## Call: survfit(formula = Surv_obj ~ 1, data = BRCA_data)
##
##       n  events  median 0.95LCL 0.95UCL
## 1027.00  104.00    9.51    8.39   12.21
##
```



The median Survival time for breast cancer patients is 9.51 Years having 95% Confidence that the median survival time will be between 8.39 Years and 12.21 Years.

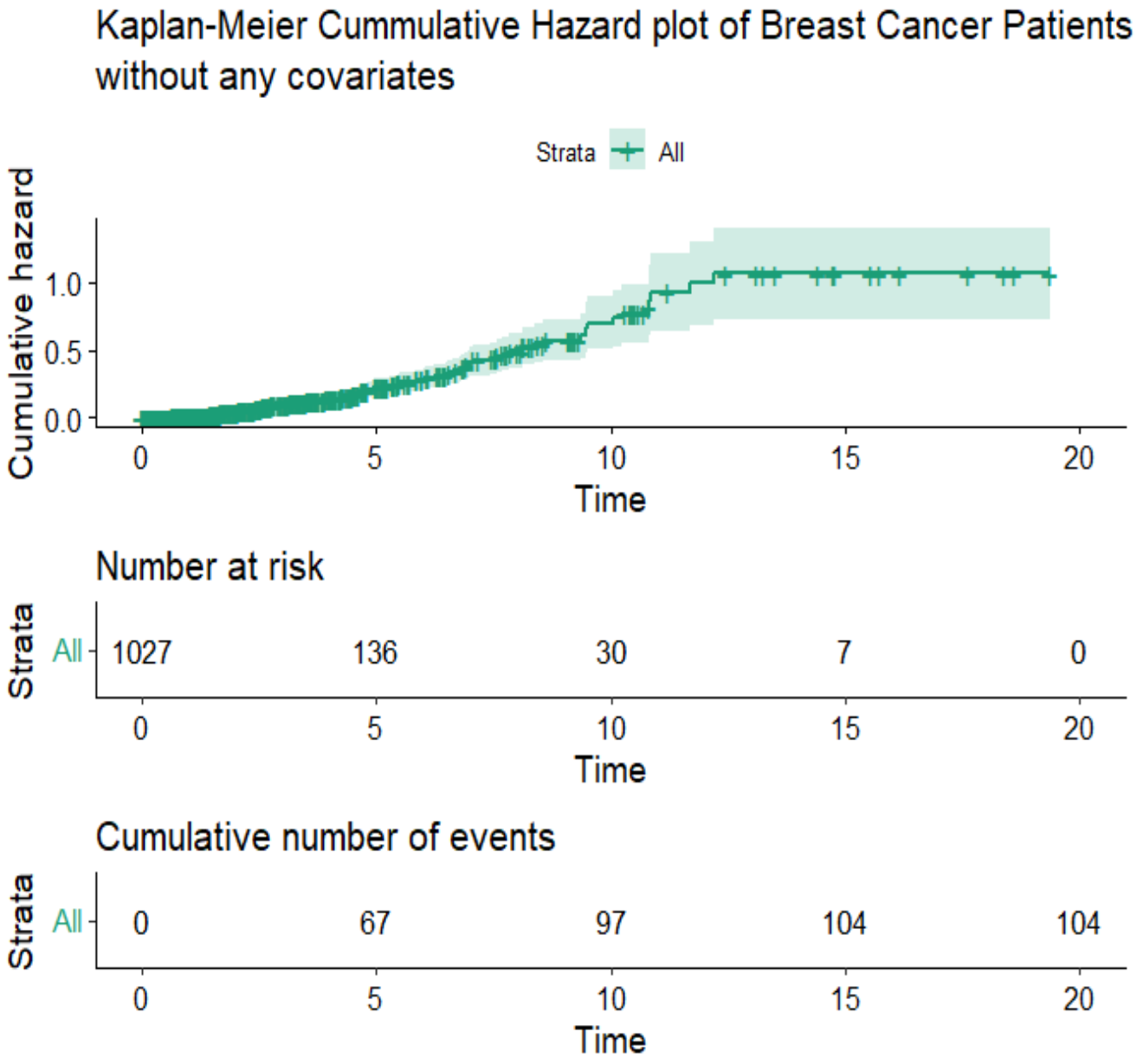
We can observe that survival decreased up to approximately 12 years then the curve became flat.

Most of the events happened in the early years.

6.2.2 Life-Table of Breast Cancer Patients

```
## Call: survfit(formula = Surv_obj ~ 1, data = BRCA_data)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.5    760       4    0.995 0.00249        0.990        1.000
##   1.0    580       9    0.981 0.00522        0.971        0.991
##   1.5    476       4    0.973 0.00652        0.960        0.986
##   2.0    407       7    0.958 0.00862        0.941        0.975
##   2.5    343      11    0.930 0.01184        0.907        0.953
##   3.0    299       9    0.904 0.01430        0.876        0.932
##   3.5    248       5    0.888 0.01577        0.857        0.919
##   4.0    206       4    0.872 0.01737        0.838        0.907
##   4.5    167       5    0.848 0.01988        0.810        0.888
##   5.0    136       9    0.798 0.02480        0.751        0.848
##   5.5    111       4    0.771 0.02737        0.719        0.827
##   6.0    100       2    0.756 0.02872        0.702        0.815
##   6.5     86       5    0.717 0.03224        0.656        0.783
##   7.0     72       6    0.664 0.03645        0.596        0.739
##   7.5     67       2    0.645 0.03775        0.575        0.723
##   8.0     55       3    0.614 0.03997        0.540        0.697
##   8.5     44       3    0.578 0.04271        0.500        0.668
##   9.0     41       1    0.564 0.04378        0.485        0.657
##   9.5     31       3    0.515 0.04845        0.428        0.619
##  10.0     30       1    0.498 0.04965        0.410        0.605
##  10.5     24       2    0.465 0.05159        0.374        0.578
##  11.0     17       3    0.396 0.05725        0.298        0.526
##  11.5     16       0    0.396 0.05725        0.298        0.526
##  12.0     15       1    0.371 0.05878        0.272        0.506
##  12.5     13       1    0.347 0.05984        0.247        0.486
##  13.0     13       0    0.347 0.05984        0.247        0.486
##  13.5     11       0    0.347 0.05984        0.247        0.486
##  14.0     10       0    0.347 0.05984        0.247        0.486
##  14.5      9       0    0.347 0.05984        0.247        0.486
##  15.0      7       0    0.347 0.05984        0.247        0.486
##  15.5      7       0    0.347 0.05984        0.247        0.486
##  16.0      5       0    0.347 0.05984        0.247        0.486
##  16.5      4       0    0.347 0.05984        0.247        0.486
##  17.0      4       0    0.347 0.05984        0.247        0.486
##  17.5      4       0    0.347 0.05984        0.247        0.486
##  18.0      3       0    0.347 0.05984        0.247        0.486
##  18.5      2       0    0.347 0.05984        0.247        0.486
##  19.0      1       0    0.347 0.05984        0.247        0.486


## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord =
pval.coord, : There are no survival curves to be compared.
##  This is a null model.
```

6.2.3



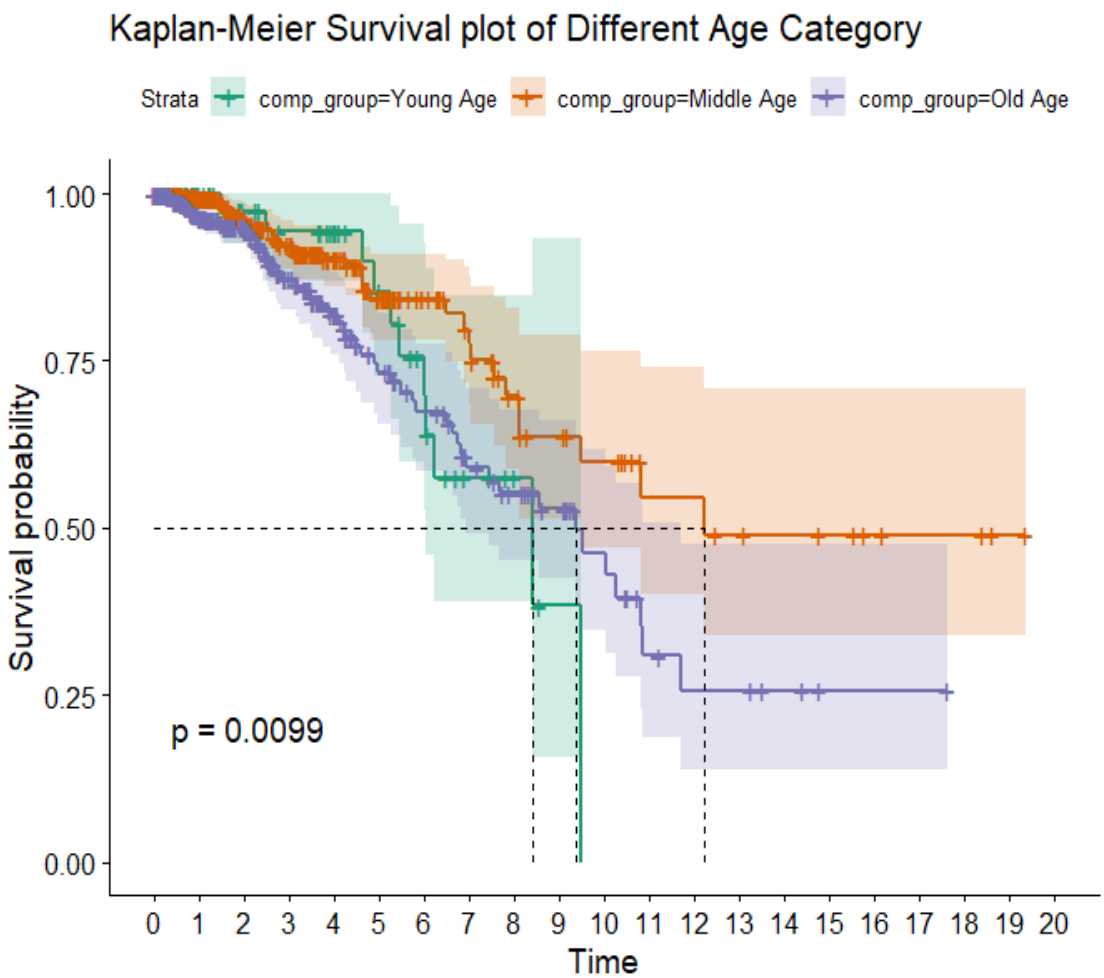
From the life table and plot, we can observe that in just 5 years Risk population became 136 from 1028 but only 67 events occurred.

This means the rest of 825 patients either had accidental death or their cancer got cured or there was a loss to follow-up.

6.3 Age

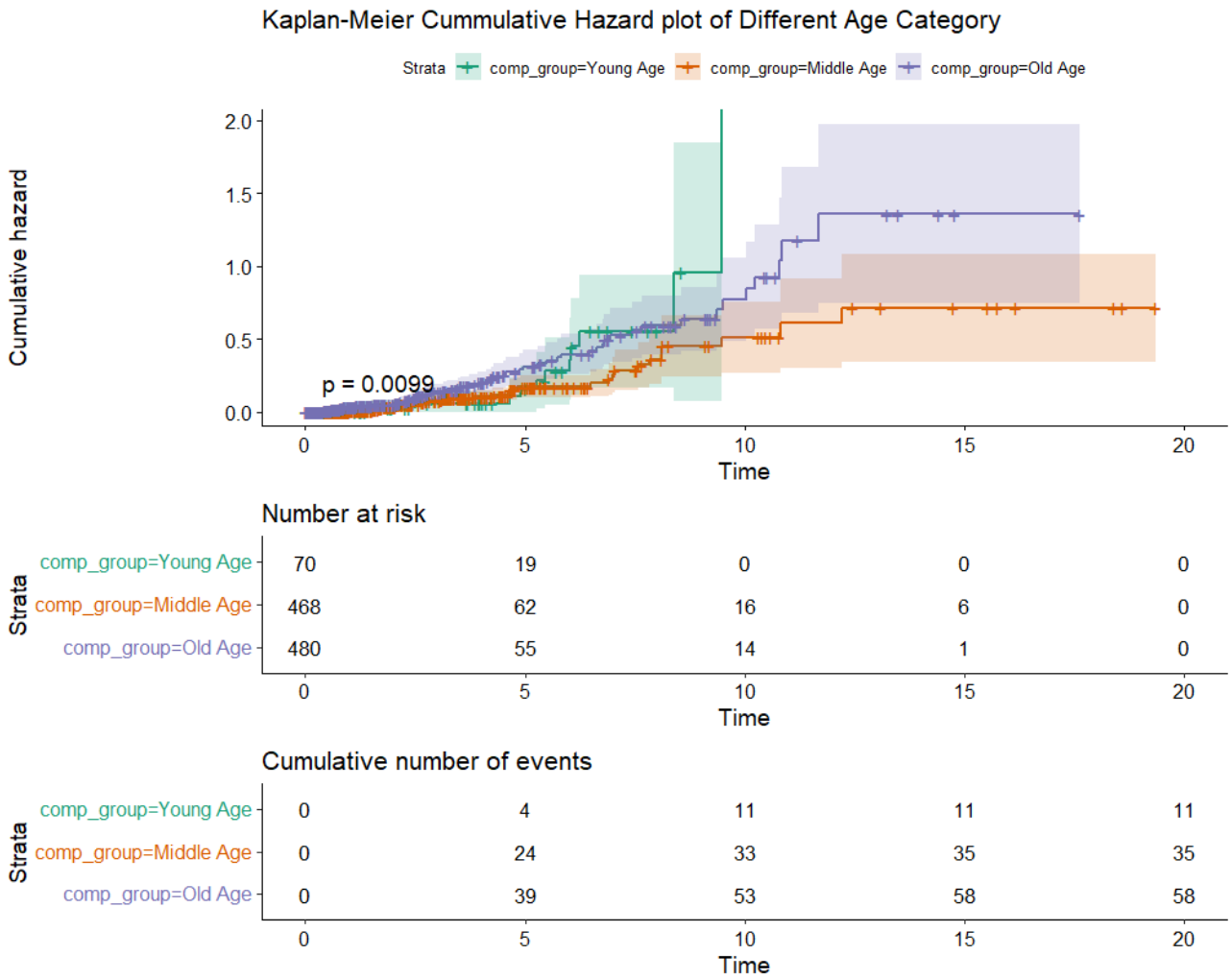
6.3.1 Non-Parametric Model fit : Kaplan-Meier

```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                          n events median 0.95LCL 0.95UCL
## comp_group=Young Age    70     11   8.39    6.05      NA
## comp_group=Middle Age 468     35  12.21    9.48      NA
## comp_group=Old Age    480     58   9.36    6.94    11.7
```



Median Survival time for the Middle-aged group is the highest, 12.21 Years with Young age having the least Median survival time, 8.39 Years, and Middle age having 9.36 Years.

Visually 3 curves seems to be different from each other.



We can observe for Patients with the Middle-aged Category have a lower hazard rate than the Young and Old age categories.
Hazard started to increase after 5 Years very rapidly for Young and for old aged category Hazard rate became steady after 12 Years.
For Middle-Aged Category Hazard rate linearly increases up to 12 Years then becomes steady.

Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Age_Category, data = temp)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## Age_Category=Young Age   70       11     9.22     0.342     0.381
## Age_Category=Middle Age 468       35    50.35     4.680     9.140
## Age_Category=Old Age    480       58    44.42     4.148     7.289
##
##  Chisq= 9.2  on 2 degrees of freedom, p= 0.01
```

We have log-rank statistic as 9.2 having chi-square distribution with p-value < 0.05. This means there is a significant difference in survival curves of different age categories.

To have the same survival, the Middle-aged category was expected to have more number events while the Young and Old aged Category was expected to have a lesser number of events.

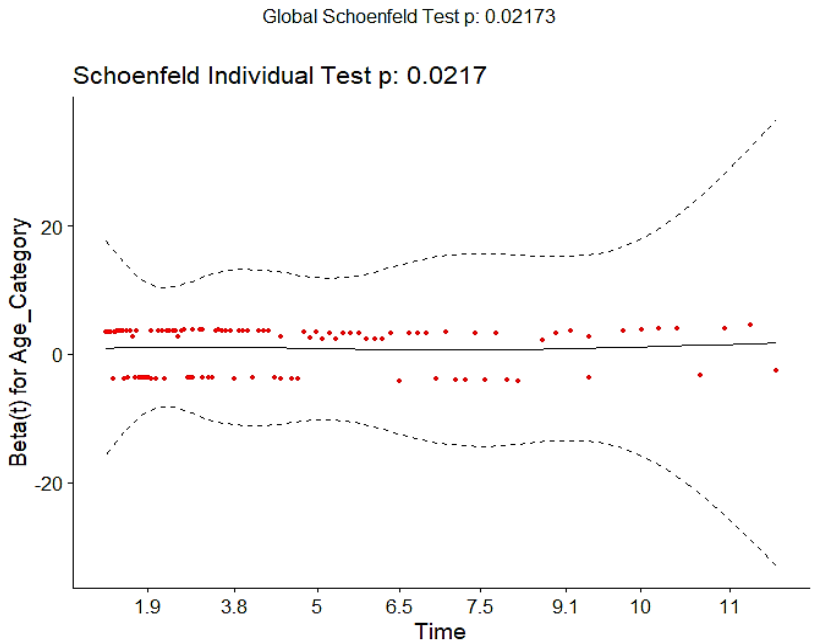
Pairwise-Difference in Survival Curves

```
##            Young Age Middle Age
## Middle Age
## Old Age              **
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

Middle and Old Aged categories have significant differece in survival curves. Rest other pairs have same survival curves.

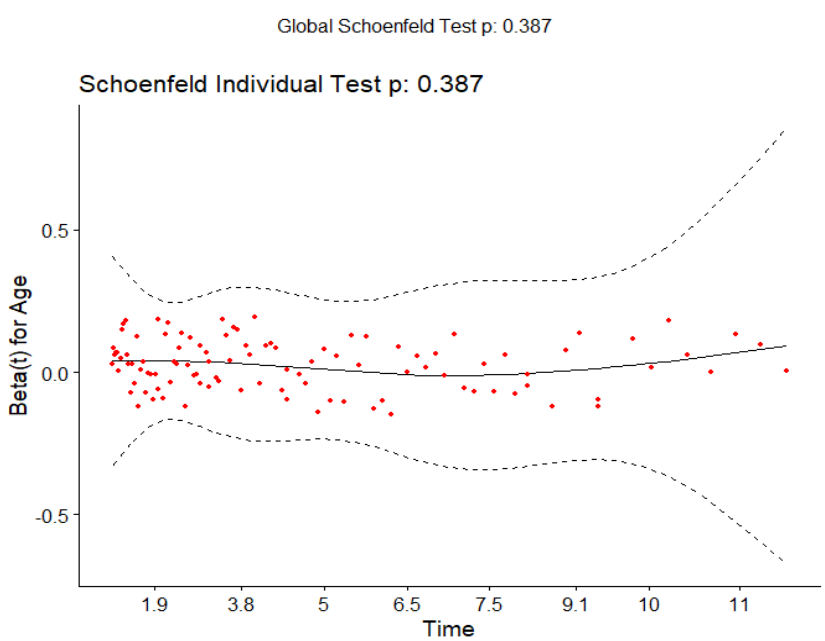
6.3.2 Semi-Parametric Model fit : Cox-Proportional Hazard Model

Proportional Hazard **Assumption** Check:



From Schoenfeld Test, our test for proportional hazard assumption suggests that Proportional hazard assumption is not met as p-value < 0.05 for Age Category co-variate. Hence, we cannot use this co-variate in the Cox-PH model.

But we can try **continuous Age Variate**.



As of now, we get P-value > 0.05 in Schoenfeld Test for proportional hazard assumption, We can use this variable in our Cox-PH model.

```
## Call:
## coxph(formula = Surv_obj ~ Age, data = temp)
##
##   n= 1018, number of events= 104
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## Age 0.026490  1.026844 0.007407 3.576 0.000348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## Age     1.027     0.9739     1.012     1.042
##
## Concordance= 0.629  (se = 0.034 )
## Likelihood ratio test= 12.75  on 1 df,   p=4e-04
## Wald test            = 12.79  on 1 df,   p=3e-04
## Score (logrank) test = 12.97  on 1 df,   p=3e-04
```

From all 3 Tests, after observing Test statistic values and P values < 0.05, we can conclude that this Cox-PH model is statistically significant.

From the hazard value i.e. exp(coef) we can conclude that for every 1 year increase in age the risk of death will increase by 2.7%

6.3.3 Parametric Model fit

```
## Distribution | AIC      ##

## Exponential  | 858.3261 ##
## Weibull      | 816.8793 ##
## Gamma        | 815.1895 ##
## Log-Normal   | 818.0849 ##
## Log-Logistic | 814.9523 ##
```

We get the lowest AIC value for **Log-Logistic Distribution**.

```
## Call:
## survreg(formula = Surv_obj ~ Age_Category, data = temp, dist = "loglogistic")
##                          Value Std. Error     z      p
## (Intercept)             2.5783     0.1238 20.83 <2e-16
## Age_CategoryYoung Age  -0.2567     0.2203 -1.17 0.2439
## Age_CategoryOld Age    -0.4254     0.1392 -3.06 0.0022
## Log(scale)             -0.6127     0.0692 -8.85 <2e-16
##
## Scale= 0.542
## Log logistic distribution
## Loglik(model)= -403.5   Loglik(intercept only)= -408.4
##  Chisq= 9.76 on 2 degrees of freedom, p= 0.0076
## Number of Newton-Raphson Iterations: 9, n= 1018
```

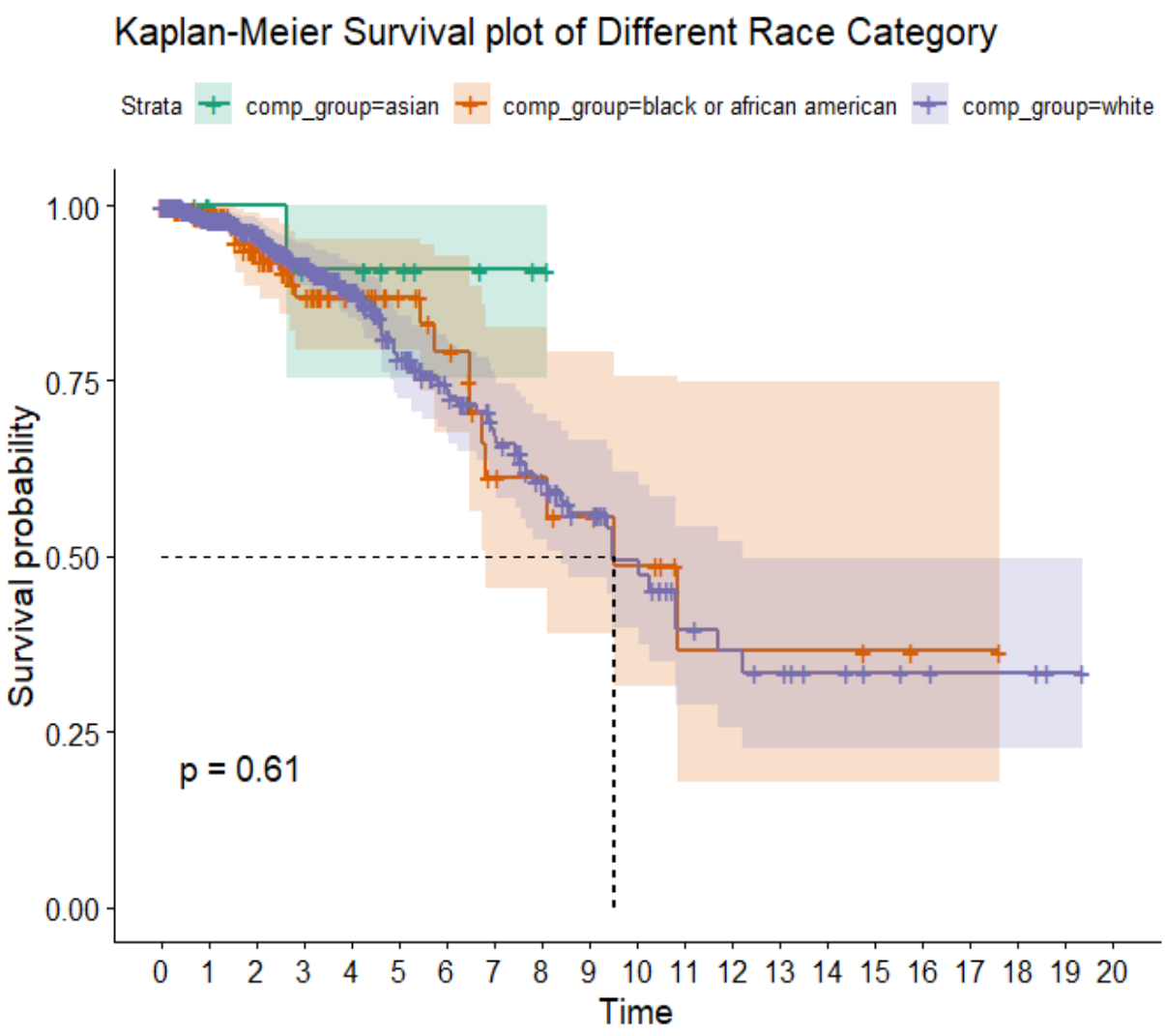
Based on Chi-Square statistic and P-value < 0.05, We can say that the Overall Model is statistically Significant.

As for the Coefficient of Old Age Category Patients, P-value < 0.05, We reject our null hypothesis that the coefficient is 0. Having ‘Old Age’ = 1 accelerates the time to event by a factor of exp(-0.4254) = 0.650 (0.650 times shorter survival time compared to the baseline survival).

6.4 Race Category

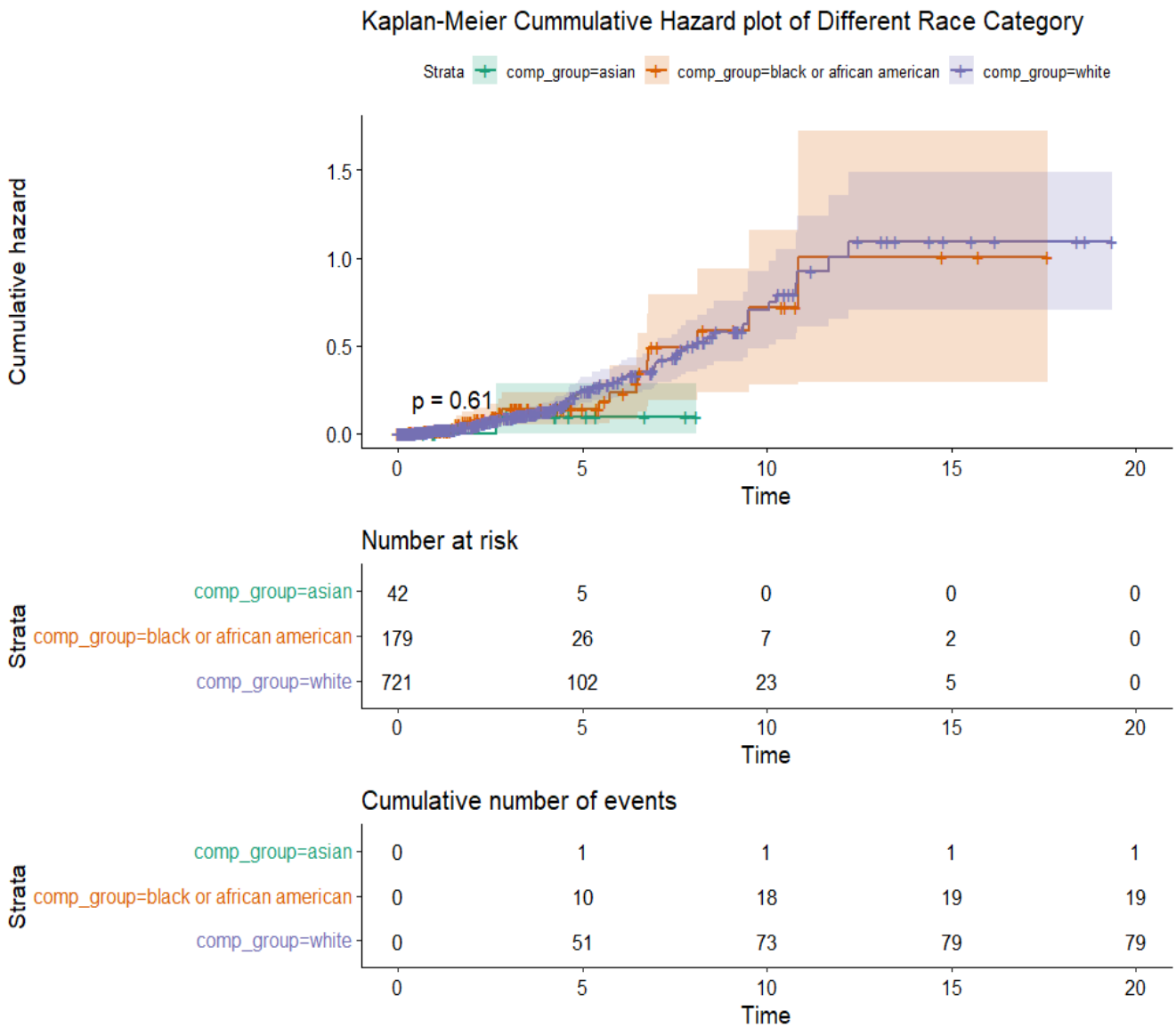
6.4.1 Non-Parametric Model fit : Kaplan-Meier

```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                                          n events median 0.95LCL 0.95UCL
## comp_group=asian                        42      1     NA      NA      NA
## comp_group=black or african american 179     19   9.51    6.80      NA
## comp_group=white                       721     79   9.48    8.39    12.2
```



We can notice that the survival curve for the Asian group didn't reach 0.50 survival probability and because of that, we cannot find Median Survival time for this group.

The median survival time for Black or African American and White group is the same around 9.50 Years.



Hazard Rate for Black or African American group increased linearly up to 11 Years then became flat and for Asian group cumulative hazard curve became flat after 12.5 years.

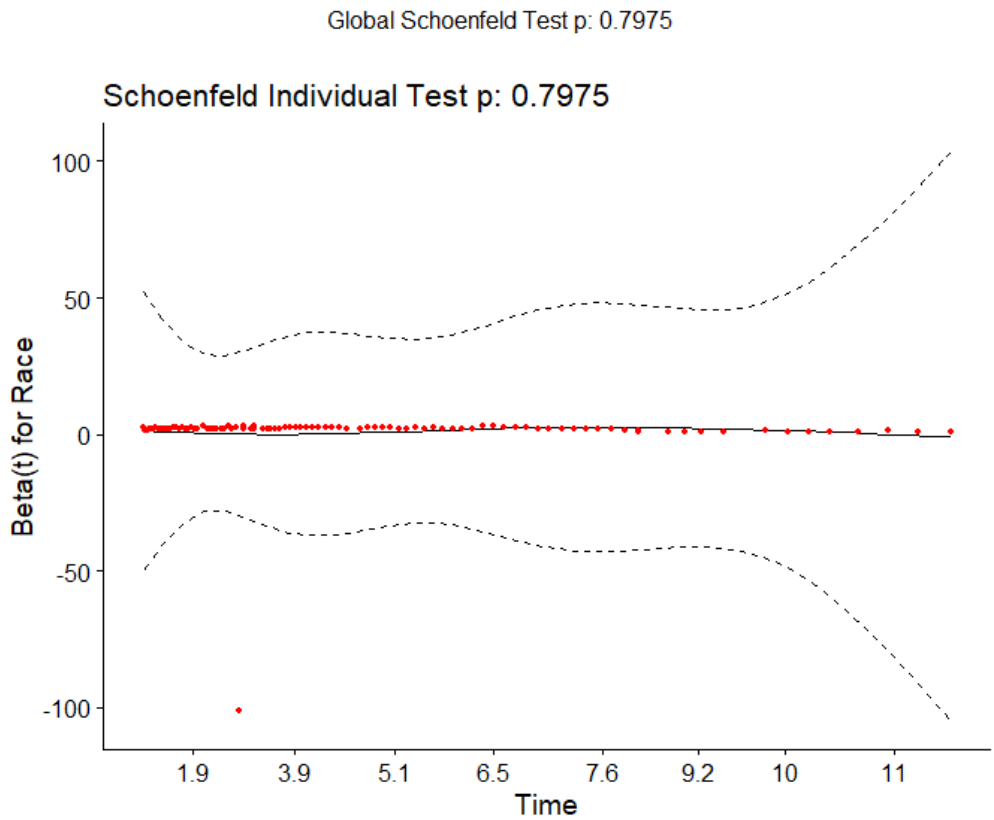
Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Race, data = temp)
##
##                                  N Observed Expected (O-E)^2/E (O-E)^2/V
## Race=asian                      42        1     2.57   0.96005    0.9918
## Race=black or african american 179       19    18.59   0.00891    0.0110
## Race=white                     721       79    77.84   0.01741    0.0817
##
##  Chisq= 1  on 2 degrees of freedom, p= 0.6

##
```

We can see that the overall log-rank statistic is equal to 1 with p-value > 0.05 This indicates that the difference in Survival Curves of different Race Categories are not statistically significant.

6.4.2 Semi-Parametric Model fit : Cox-Proportional Hazard Model

Proportional Hazard **Assumption** Check:



As Schoenfeld residuals are randomly distributed around the mean and our p value > 0.05 in Schoenfeld's Test, Our assumption of Proportional hazard is met for Race covariate.

```
## Call:
## coxph(formula = Surv_obj ~ Race, data = temp)
##
##   n= 942, number of events= 99
##
##                                 coef exp(coef) se(coef)     z Pr(>|z|)
## Raceblack or african american 0.9703    2.6387   1.0274 0.944    0.345
## Racewhite                     0.9629    2.6194   1.0074 0.956    0.339
##
##                               exp(coef) exp(-coef) lower .95 upper .95
## Raceblack or african american     2.639     0.3790    0.3522     19.77
## Racewhite                         2.619     0.3818    0.3637     18.87
##
## Concordance= 0.517  (se = 0.027 )
## Likelihood ratio test= 1.29  on 2 df,   p=0.5
## Wald test            = 0.92  on 2 df,   p=0.6
## Score (logrank) test = 0.99  on 2 df,   p=0.6
```

From all three-test statistics, we get p value > 0.05 which indicates that coefficients of Different race categories are not significantly different from 0. Hence, Race Category is not a good predictor of hazard.

6.3.3 Parametric Model fit

```
## Distribution | AIC      ##

## Exponential  | 824.6335 ##
## Gamma        | 785.5206 ##
## Log-Normal   | 790.1024 ##
## Log-Logistic | 784.3595 ##
```

We get the lowest AIC value for **Log-Logistic Distribution**.

```
##
## Call:
## survreg(formula = Surv_obj ~ Race, data = temp, dist = "loglogistic")
##                                  Value Std. Error     z       p
## (Intercept)                     2.9197     0.5843  5.00 5.8e-07
## Raceblack or african american -0.5911     0.5981 -0.99    0.32
## Racewhite                     -0.5616     0.5831 -0.96    0.34
## Log(scale)                    -0.6086     0.0719 -8.47 < 2e-16
##
## Scale= 0.544
##
## Log logistic distribution
## Loglik(model)= -388.2   Loglik(intercept only)= -388.8
##  Chisq= 1.27 on 2 degrees of freedom, p= 0.53
## Number of Newton-Raphson Iterations: 7
## n= 942
```

Based on Chi-Square statistic and P-value > 0.05, We can say that Overall Model is statistically insignificant.

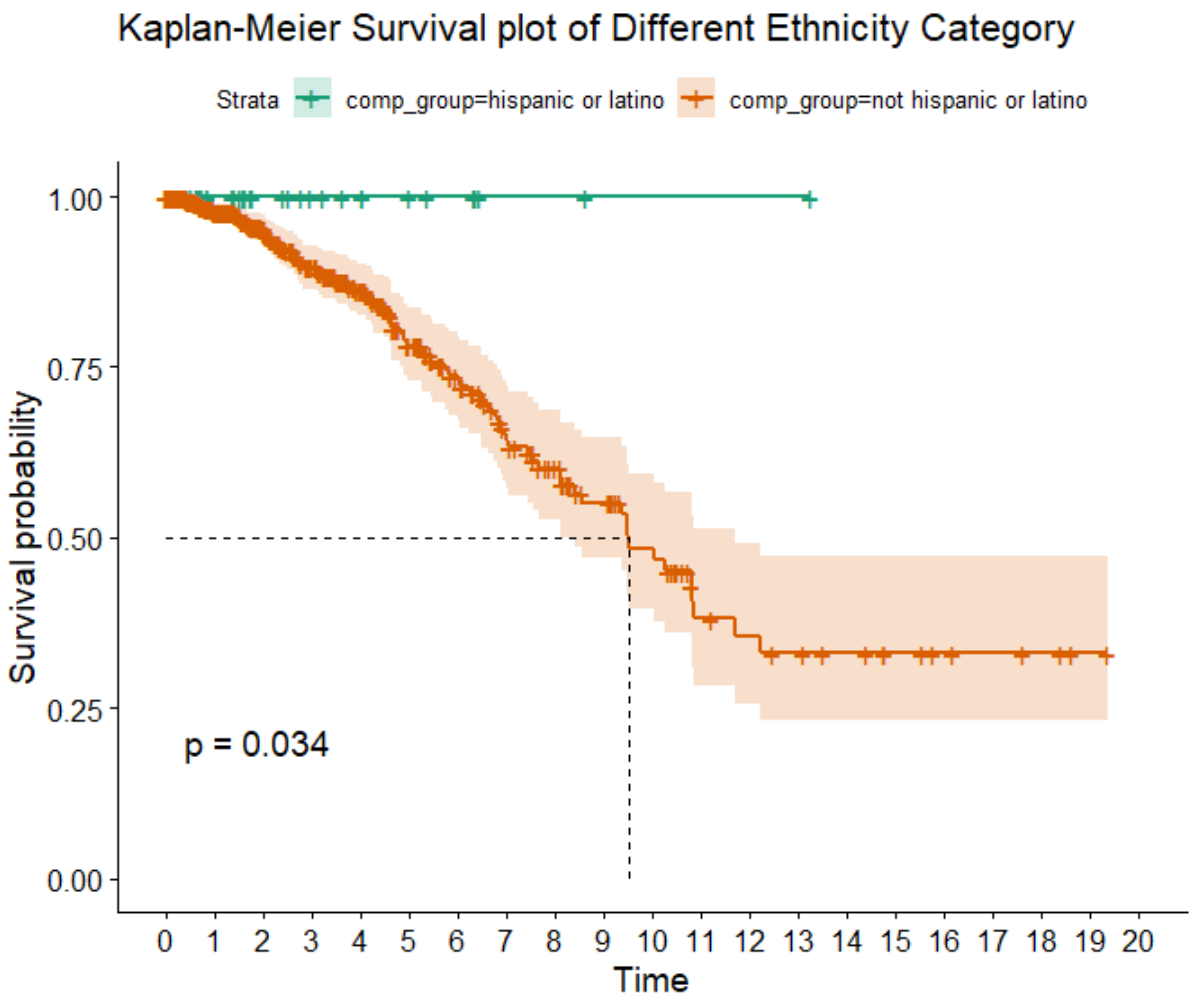
Hence, we cannot use Race Variable in Parametric Modeling of Breast Cancer Patients.

6.5 Ethnicity Category

The proportion of Ethnicity groups are very unbalanced, Hispanic or Latino groups have very small risk population as well as we have 0 number of events in this group.

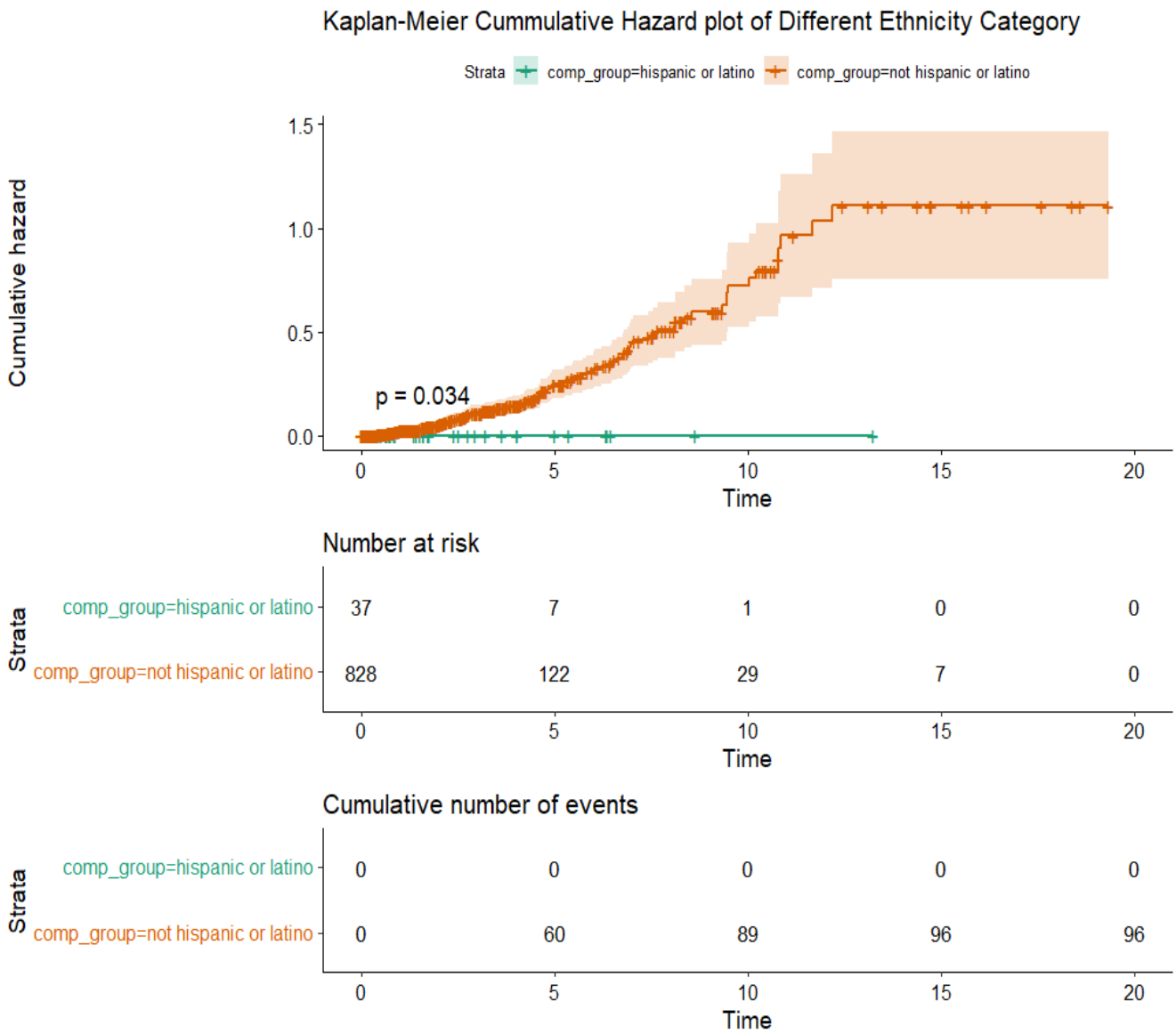
6.5.1 Non-Parametric Model fit: Kaplan-Meier

```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                                      n events median 0.95LCL 0.95UCL
## comp_group=hispanic or latino       37      0     NA      NA      NA
## comp_group=not hispanic or latino 828     96   9.51    8.39    11.7
```



We can Observe clearly that the Hispanic or Latino group has a flat survival curve because no event was reported in the 20 Years.

The median survival time for Not Hispanic or Latino group is 9.51 Years while we KM-Model failed to calculated Median Survival time for the Hispanic or Latino group as the survival curve didn't reach the Survival probability of 0.50



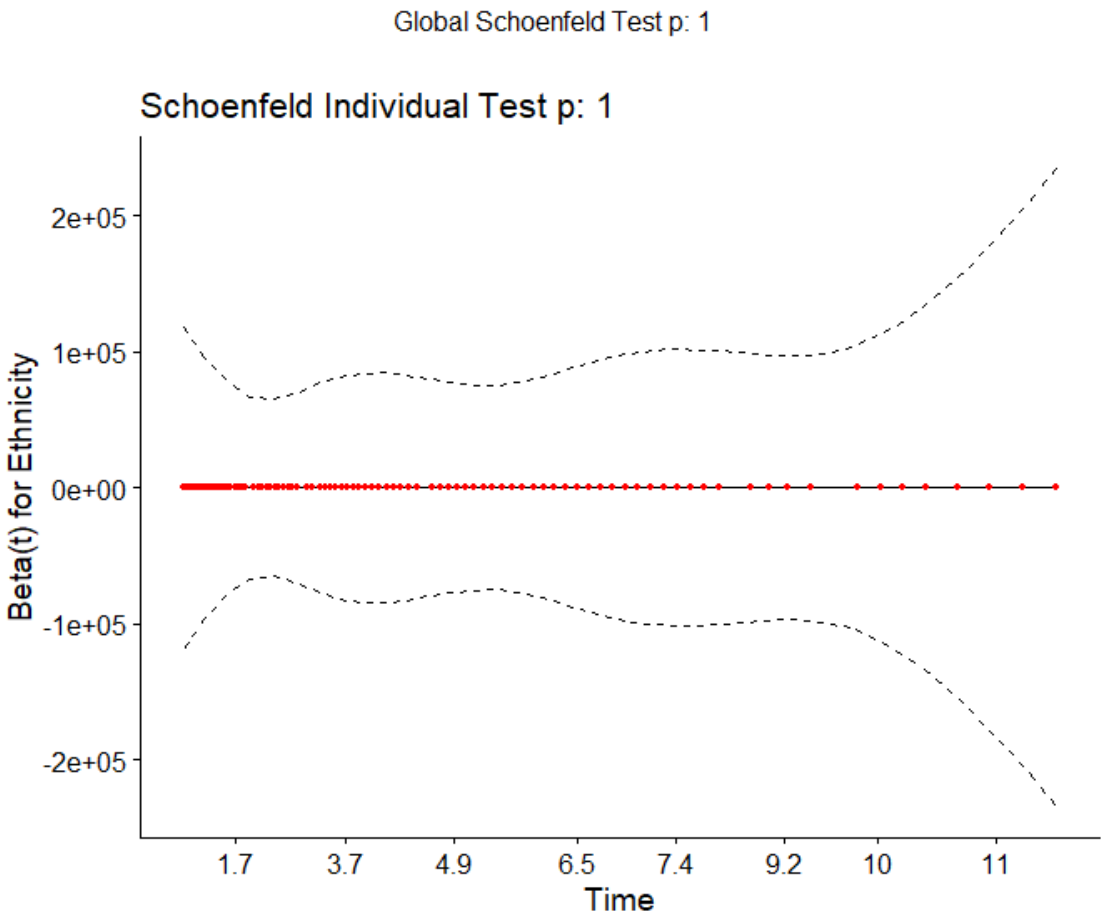
The hazard rate of Not Hispanic or Latino groups increased linearly up to 12 Years and then became steady whereas the Hazard rate for the Hispanic or Latino group was steady all the time with no significant changes.

Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Ethnicity, data = temp)
##
##                                    N Observed Expected (O-E)^2/E (O-E)^2/V
## Ethnicity=hispanic or latino      37        0     4.27     4.272      4.48
## Ethnicity=not hispanic or latino 828       96    91.73     0.199      4.48
##
##  Chisq= 4.5  on 1 degrees of freedom, p= 0.03
```

From log-rank Test Statistic 4.5 following chi-square distribution with 1 degree of freedom and p-value < 0.05 we can conclude that there is a significant difference in survival curves of both Ethnicity groups.

6.3.2 Semi-Parametric Model fit : Cox-Proportional Hazard Model




```
## Call:
## coxph(formula = Surv_obj ~ Ethnicity, data = temp)
##
##   n= 865, number of events= 96
##
##                                      coef exp(coef)  se(coef)     z Pr(>|z|)
## Ethnicitynot hispanic or latino 1.708e+01 2.604e+07 2.410e+03 0.007    0.994
##
##                                 exp(coef) exp(-coef) lower .95 upper .95
## Ethnicitynot hispanic or latino  26041085   3.84e-08         0       Inf
##
## Concordance= 0.524  (se = 0.005 )
## Likelihood ratio test= 8.75  on 1 df,   p=0.003
## Wald test            = 0  on 1 df,   p=1
## Score (logrank) test = 4.48  on 1 df,   p=0.03

## Warning in fitter(X, Y, istrat, offset, init, control, weights = weights, :
## Loglik converged before variable 1 ; coefficient may be infinite.
```

When we tried using the Ethnicity variable as a covariate in the Cox-PH model then Log-Likelihood converged before co-variate and we got an infinite coefficient. Hence we cannot we this variable to predict the hazard of breast cancer patients.

6.3.3 Parametric Model fit

```
## Distribution | AIC      ##

## Exponential  | 779.2353 ##
## Weibull      | 747.2569 ##
## Gamma        | 745.4185 ##
## Log-Normal   | 749.0763 ##
## Log-Logistic | 744.187 ##
```

We get the lowest AIC value for **Log-Logistic Distribution.**

```
##
## Call:
## survreg(formula = Surv_obj ~ Ethnicity, data = temp, dist = "loglogistic")
##                                    Value Std. Error        z       p
## (Intercept)                     5.50e+03   8.60e-02 63953.85 < 2e-16
## Ethnicitynot hispanic or latino -5.50e+03   0.00e+00     -Inf < 2e-16
## Log(scale)                      -5.89e-01   7.35e-02    -8.01 1.1e-15
##
## Scale= 0.555
##
## Log logistic distribution
## Loglik(model)= -369.1   Loglik(intercept only)= -373.4
##  Chisq= 8.63 on 1 degrees of freedom, p= 0.0033
## Number of Newton-Raphson Iterations: 8
## n= 865
```

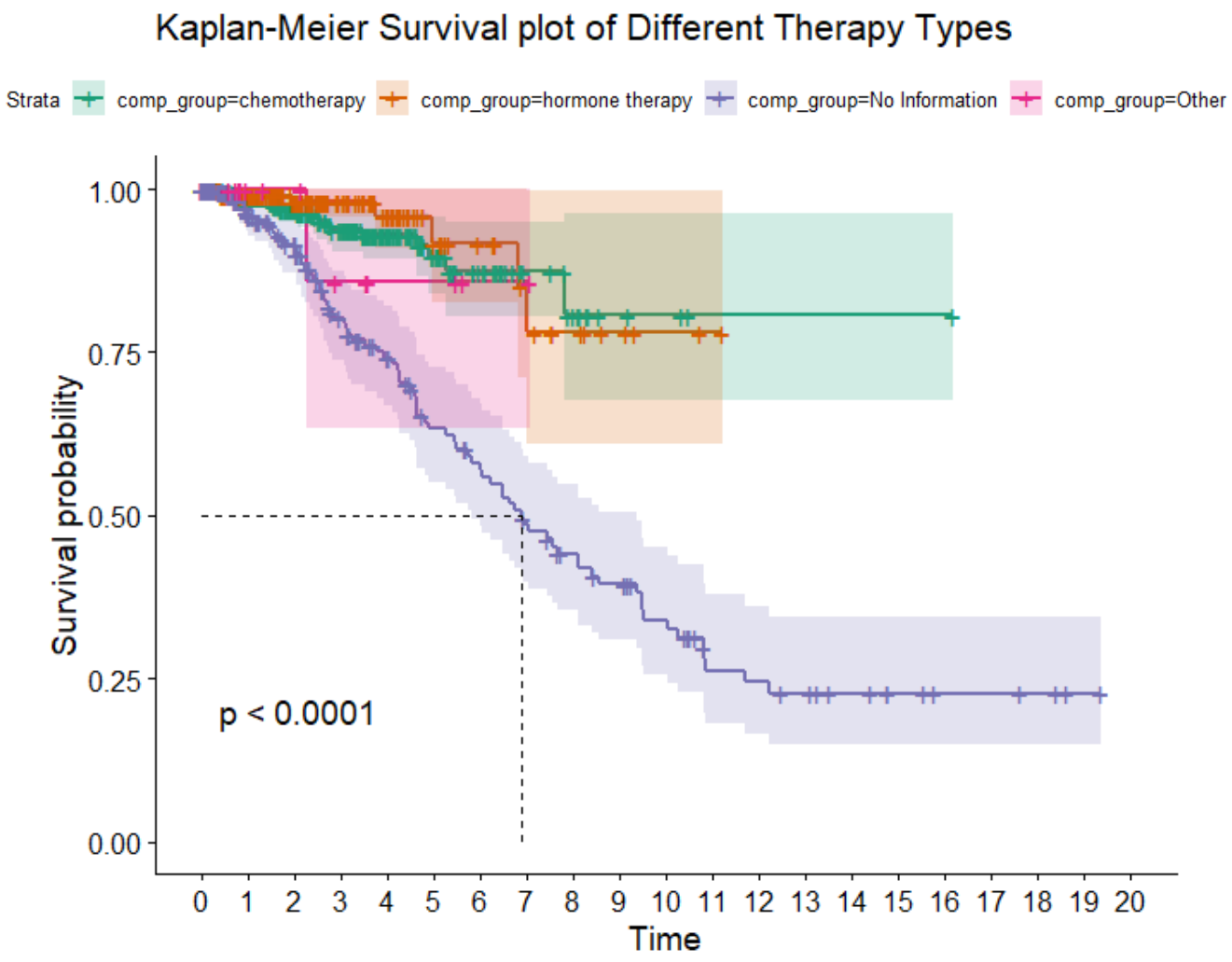
Based on Chi-Square statistic and P-value < 0.05, the Overall Model seems statistically Significant but, Having 'not hispanic or latino' = 1 accelerates the time to event by a factor of exp(-5.50e+03) = 0 (0 times shorter survival time compared to the baseline survival).

Hence, We cannot use the Ethnicity variable to predict Survival for Breast Cancer Patients.

6.6 Therapy Type Category

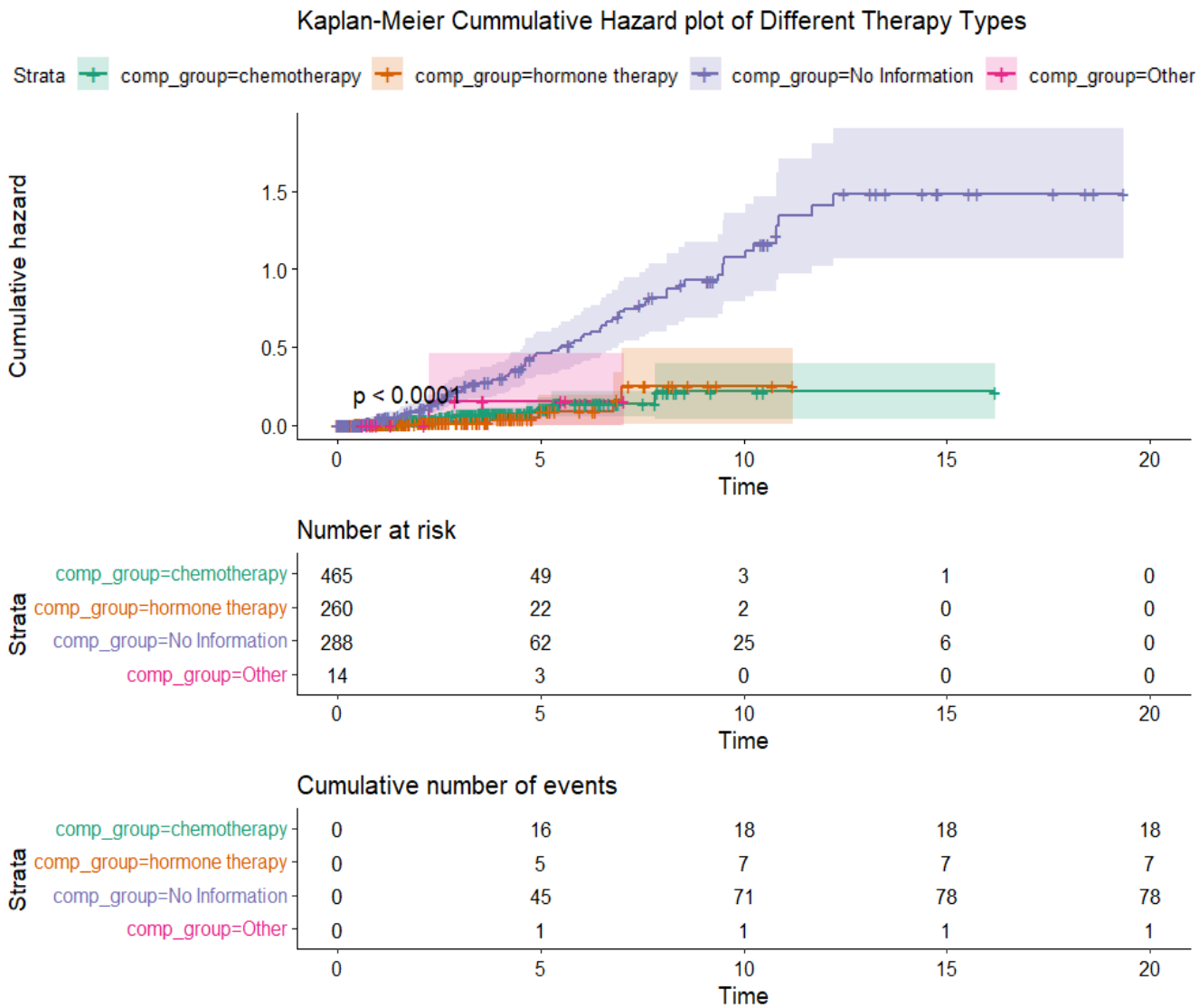
6.6.1 Non-Parametric Model fit : Kaplan-Meier

```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                                n events median 0.95LCL 0.95UCL
## comp_group=chemotherapy    465     18     NA      NA      NA
## comp_group=hormone therapy 260      7     NA      NA      NA
## comp_group=No Information  288     78    6.9    5.83    9.36
## comp_group=Other            14      1     NA      NA      NA
```



Median Survival time is 6.9 Years for patients for which No Therapy type is reported.

For Hormone Therapy, Chemotherapy and Other therapies Survival curve didn't reach Survival probability of 0.50 and because of that KM-Model failed to compute Median Survival times for these Therapy groups.



Visually we can't see any difference in hazard rates for patients who received any of the Breast cancer therapies.

Hazard Rate of Breast Cancer patients with No therapy type reported increased linearly upto 12.5 Years after first diagnosis then became steady.

Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Therapy_Type, data = temp)
##
##                                N Observed Expected (O-E)^2/E (O-E)^2/V
## Therapy_Type=chemotherapy    465       18    36.35     9.260    15.492
## Therapy_Type=hormone therapy 260        7    18.97     7.556     9.491
## Therapy_Type=No Information  288       78    47.20    20.107    43.891
## Therapy_Type=Other            14        1     1.49     0.159     0.162
##
##  Chisq= 44.1  on 3 degrees of freedom, p= 1e-09
```

We got large Log-rank statistic value, 44.1 following Chi-Square distribution with 3 degrees of freedom and p value < 0.05, This indicates significant difference in survival curves of Different Therapy types.

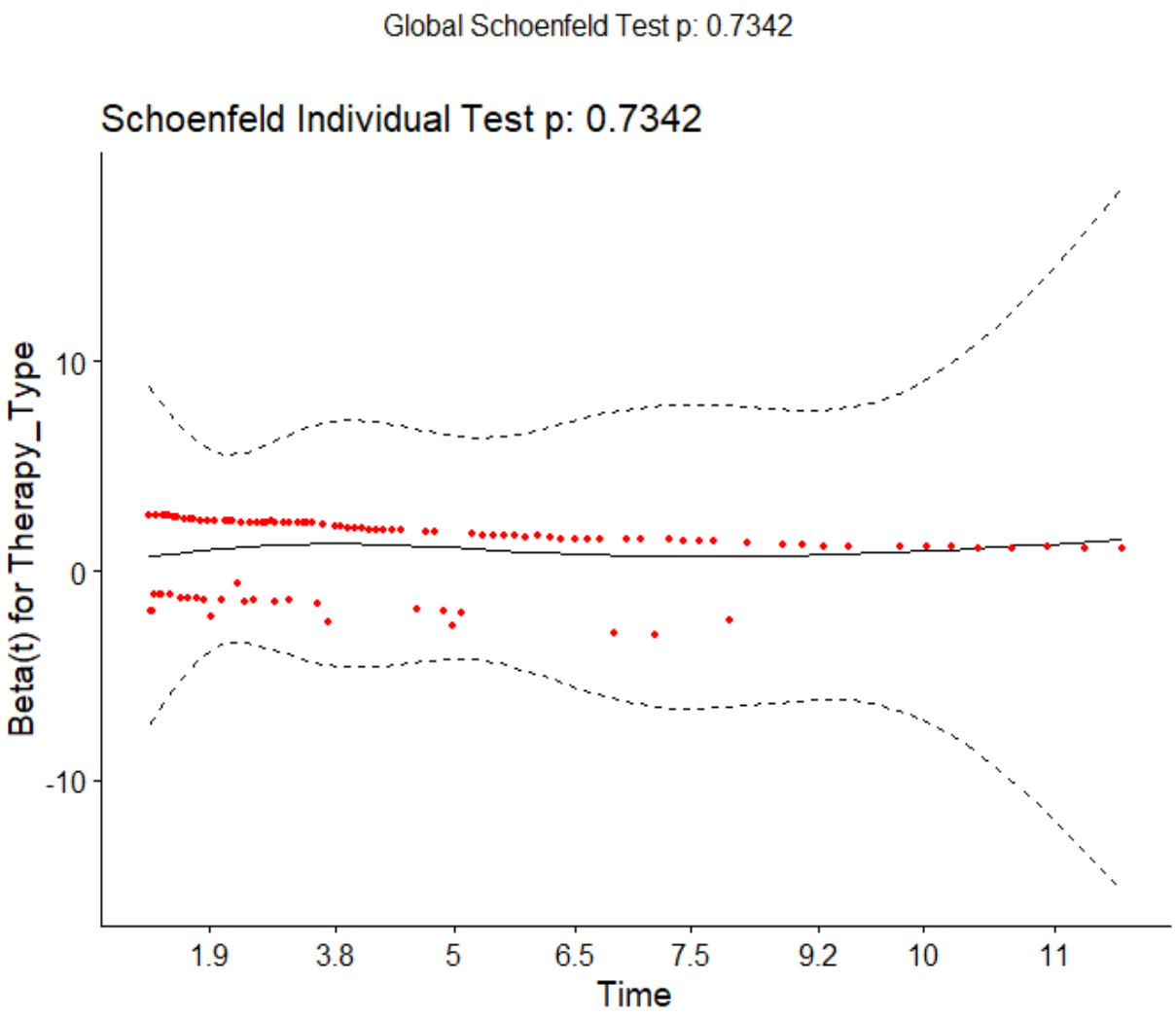
Pairwise-Difference in Survival Curves

```
##                  chemotherapy hormone therapy No Information
## hormone therapy
## No Information  ****         ****
## Other
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

We can see there is significant difference in patients having Chemotherapy and Patients with No therapy reported. Same with patients having Hormone Therapy and with No therapy.

6.6.2 Semi-Parametric Model fit : Cox-Proportional Hazard Model

Proportional Hazard **Assumption** Check:



Schoenfeld's Residuals are randomly distributed distributed around mean 0 and p value > 0.05 and because of that we'll fail to reject our null hypothesis of proportional hazard. Therefore, Proportional hazard assumption is met.

```
## Call:
## coxph(formula = Surv_obj ~ Therapy_Type, data = temp)
##
##   n= 1027, number of events= 104
##
##                                coef exp(coef) se(coef)      z Pr(>|z|)
## Therapy_Typehormone therapy -0.2835    0.7532   0.4456 -0.636    0.525
## Therapy_TypeNo Information   1.3880    4.0068   0.2704  5.132 2.86e-07 ***
## Therapy_TypeOther            0.3024    1.3531   1.0277  0.294    0.769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                             exp(coef) exp(-coef) lower .95 upper .95
## Therapy_Typehormone therapy    0.7532     1.3277    0.3145     1.804
## Therapy_TypeNo Information     4.0068     0.2496    2.3583     6.808
## Therapy_TypeOther              1.3531     0.7391    0.1805    10.141
##
## Concordance= 0.672  (se = 0.033 )
## Likelihood ratio test= 44.28  on 3 df,   p=1e-09
## Wald test            = 38.3  on 3 df,   p=2e-08
## Score (logrank) test = 44.09  on 3 df,   p=1e-09
```

We can see the coefficient 1.388 of No Information group is statistically significant with P-value < 0.05. Cumulative Hazard rate exp(coef) for no information group is 4 times larger than chemotherapy or we can say approximately 4 times larger than the patients having any treatment as all treatments leads to insignificantly different survival curves for all patient having any Breast cancer treatment.

6.6.3 Parametric Model fit

```
## Distribution | AIC      ##

## Exponential  | 805.6127 ##
## Gamma        | 786.4540 ##
## Log-Normal   | 789.8137 ##
## Log-Logistic | 782.8935 ##
```

We get lowest AIC value for **Log-Logistic Distribution.**

```
## Call:
## survreg(formula = Surv_obj ~ Therapy_Type, data = temp, dist = "loglogistic")
##                               Value Std. Error     z       p
## (Intercept)                  1.9553     0.0968 20.20 < 2e-16
## Therapy_Typechemotherapy     0.9309     0.1941  4.80 1.6e-06
## Therapy_Typehormone therapy  1.1203     0.2701  4.15 3.4e-05
## Therapy_TypeOther            0.6971     0.6480  1.08    0.28
## Log(scale)                  -0.5250     0.0738 -7.11 1.1e-12
##
## Scale= 0.592
##
## Log logistic distribution
## Loglik(model)= -386.4   Loglik(intercept only)= -408.8
##  Chisq= 44.73 on 3 degrees of freedom, p= 1.1e-09
## Number of Newton-Raphson Iterations: 9
## n= 1027
```

Took “No information” (No Therapy type reported) as baseline survival for comparison.

Based on Chi-Square statistic and P-value < 0.05, We can say that Overall Model is statistically Significant.

As for Coefficient of Chemotherapy and Hormone Therapy treated Patients, P-value < 0.05, We reject our null hypothesis that coefficient is 0. But for Other Therapies we fail to reject our null Hypothesis.

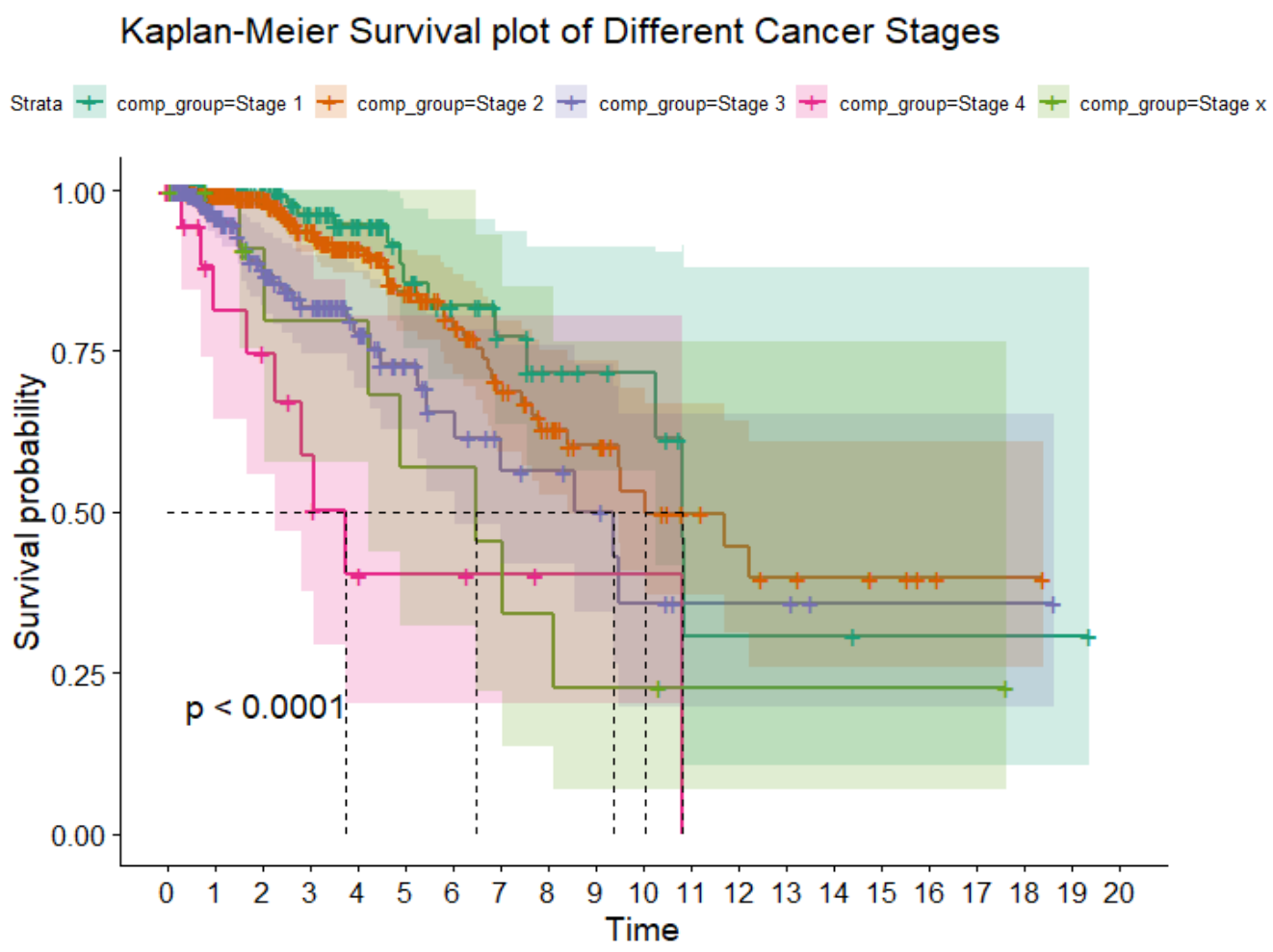
Having ‘Chemotherapy’ = 1 accelerates the time to event by a factor of exp(0.9309) = 2.54 (2.54 times longer survival time compared to the baseline survival).

Having ‘Hormone Therapy’ = 1 accelerates the time to event by a factor of exp(1.1203) = 3.06 (3.06 times longer survival time compared to the baseline survival).

6.7 Cancer Stage

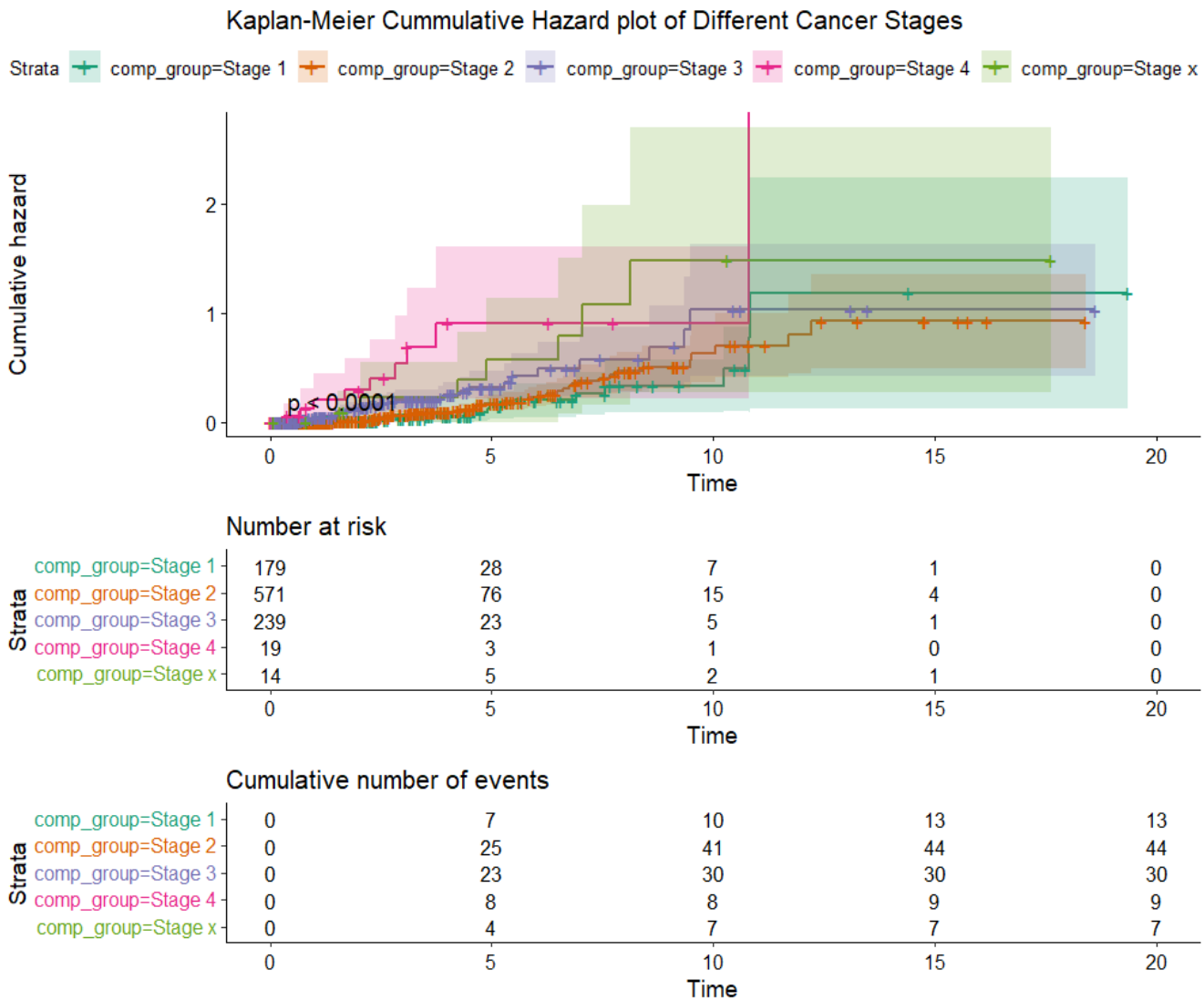
6.7.1 Non-Parametric Model fit : Kaplan-Meier

```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                        n events median 0.95LCL 0.95UCL
## comp_group=Stage 1 179     13  10.81   10.24      NA
## comp_group=Stage 2 571     44  10.05    8.39      NA
## comp_group=Stage 3 239     30   9.36    6.05      NA
## comp_group=Stage 4  19      9   3.74    2.26      NA
## comp_group=Stage x  14      7   6.50    4.22      NA
```



We can notice as Cancer Stage increases Survival probability decreases with time. Median Survival time of Patients having Stage 1 or Stage 2 Breast cancers are almost equal around 10.5 Years.

Median Survival time for Cancer Stage 4 is 3.74 which is least in different Cancer Stage groups and Cancer Stage Stage x when Cancer stage cannot be evaluated, Median survival time is 6.50 which is comparatively lower than Cancer Stages 1,2 and 3.



For Stage 4 Breast Cancer Hazard Rate increased very quickly from early years after first diagnosis and for other Cancer Stages Hazard rate increased fastly after 5 Years after first diagnosis.

Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Cancer_Stage, data = temp)
##
##                          N Observed Expected (O-E)^2/E (O-E)^2/V
## Cancer_Stage=Stage 1 179       13    21.61      3.43      4.35
## Cancer_Stage=Stage 2 571       44    55.96      2.56      5.61
## Cancer_Stage=Stage 3 239       30    19.44      5.74      7.10
## Cancer_Stage=Stage 4  19        9     2.42     17.87     18.36
## Cancer_Stage=Stage x  14        7     3.58      3.28      3.43
##
##  Chisq= 33  on 4 degrees of freedom, p= 1e-06
```

From high Log rank statistic value 33 which follows chi-square distribution with 4 degrees of freedom and P-value < 0.05 we can interpret that there is significant difference in survival curves of different Cancer Stages.

Pairwise-Difference in Survival Curves

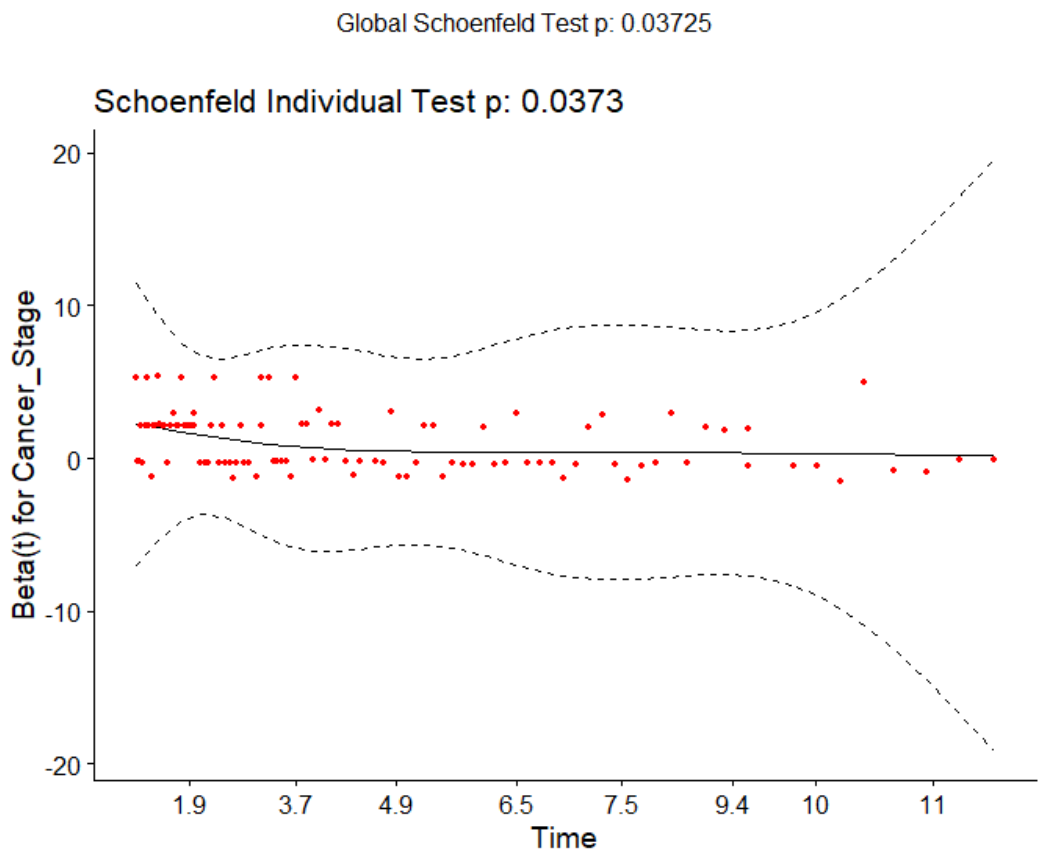
```
##         Stage 1 Stage 2 Stage 3 Stage 4
## Stage 2
## Stage 3 *       *
## Stage 4 ****    ****    *
## Stage x *       +
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

Stage 1 survival curve is significantly different from Stage 3,4 and Stage x. Stage 2 survival curve is significantly different from Stage 3,4.

Stage 3 survival curve is significantly different from stage 1,2,4. Stage 4 survival curve is significantly different from stage 1,2,3.

6.7.2 Semi-Parametric Model fit : Cox-Proportional Hazard Model

Proportional Hazard **Assumption** Check:



From Schoenfeld Test we get p-value < 0.05 which rejects our null hypothesis of proportional hazard assumption.

Hence, we cannot use Therapy Type to Predict Hazard of Breast cancer patients using Cox-PH model.

6.7.3 Parametric Model fit

```
## Distribution | AIC      ##

## Exponential  | 840.3717 ##
## Weibull      | 802.3779 ##
## Gamma        | 796.7449 ##
## Log-Normal   | 792.7795 ##
## Log-Logistic | 792.4570 ##
```

We get lowest AIC value for **Log-Logistic Distribution**.

```
##
## Call:
## survreg(formula = Surv_obj ~ Cancer_Stage, data = temp, dist = "loglogistic")
##                       Value Std. Error     z      p
## (Intercept)          2.6834     0.1790 14.99 <2e-16
## Cancer_StageStage 2 -0.1862     0.1903 -0.98 0.3280
## Cancer_StageStage 3 -0.6761     0.2064 -3.28 0.0011
## Cancer_StageStage 4 -1.4122     0.3162 -4.47  8e-06
## Cancer_StageStage x -0.8604     0.3443 -2.50 0.0124
## Log(scale)          -0.6371     0.0703 -9.06 <2e-16
##
## Scale= 0.529
##
## Log logistic distribution
## Loglik(model)= -390.2   Loglik(intercept only)= -405.9
##  Chisq= 31.28 on 4 degrees of freedom, p= 2.7e-06
## Number of Newton-Raphson Iterations: 9
## n= 1022
```

Took “Stage 1” as baseline survival for comparison.

Based on Chi-Square statistic and P-value < 0.05, We can say that Overall Model is statistically Significant and at least one coefficient is significantly different from 0.

As for Coefficient of Stage 3, Stage 4, and Stage x Cancer Patients, P-value < 0.05, We reject our null hypothesis that coefficient is 0. But for Stage 2 Cancer we fail to reject our null Hypothesis.

Having ‘Stage 3’ = 1 accelerates the time to event by a factor of exp(-0.6761) = 0.50 (0.50 times shorter survival time compared to the baseline survival).

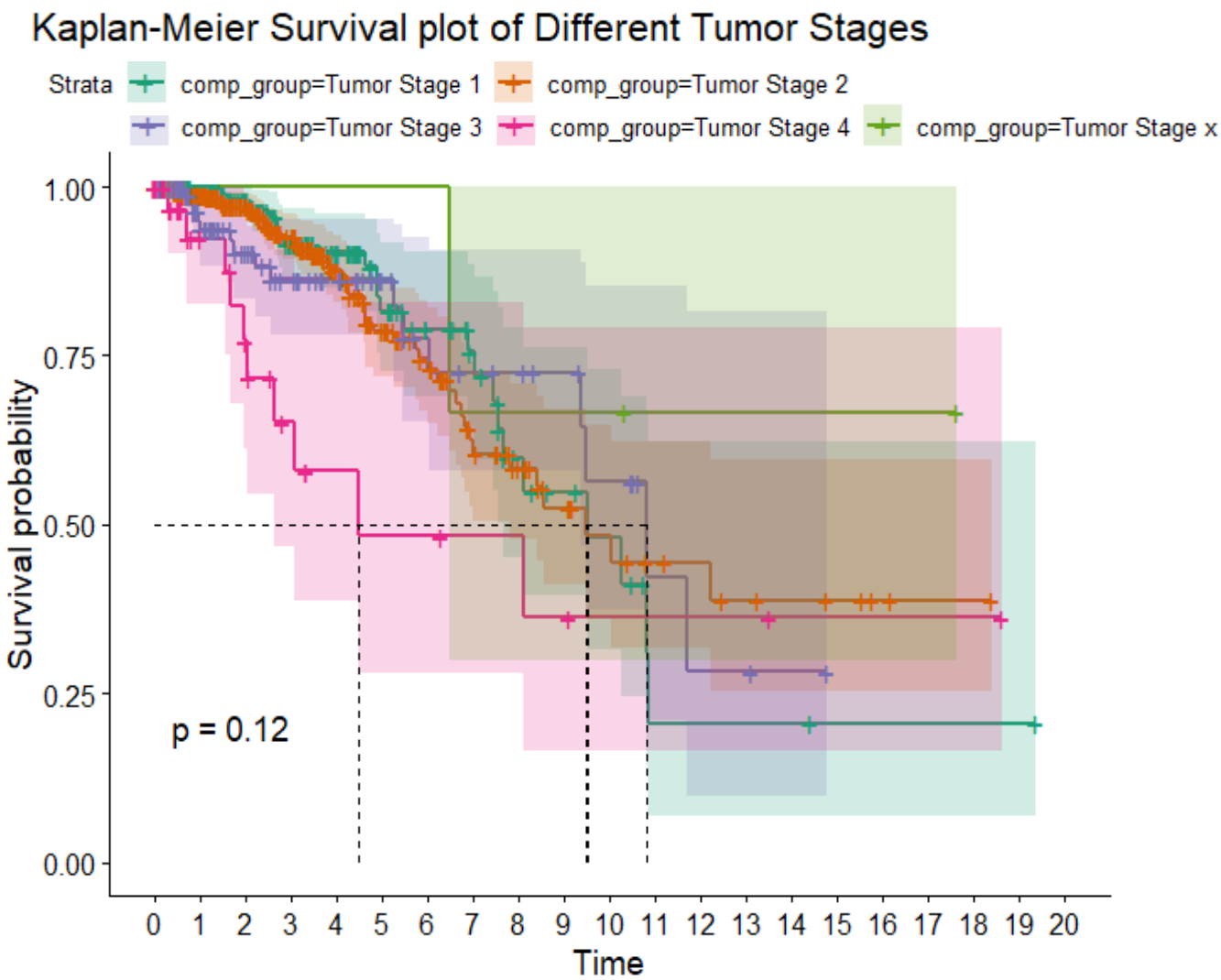
Having ‘Stage 4’ = 1 accelerates the time to event by a factor of exp(-1.4122) = 0.24 (0.24 times shorter survival time compared to the baseline survival).

Having ‘Stage x’ = 1 (When Cancer Stage cannot be assessed) accelerates the time to event by a factor of exp(-0.8604) = 0.42 (0.42 times shorter survival time compared to the baseline survival).

6.8 Tumor Stage

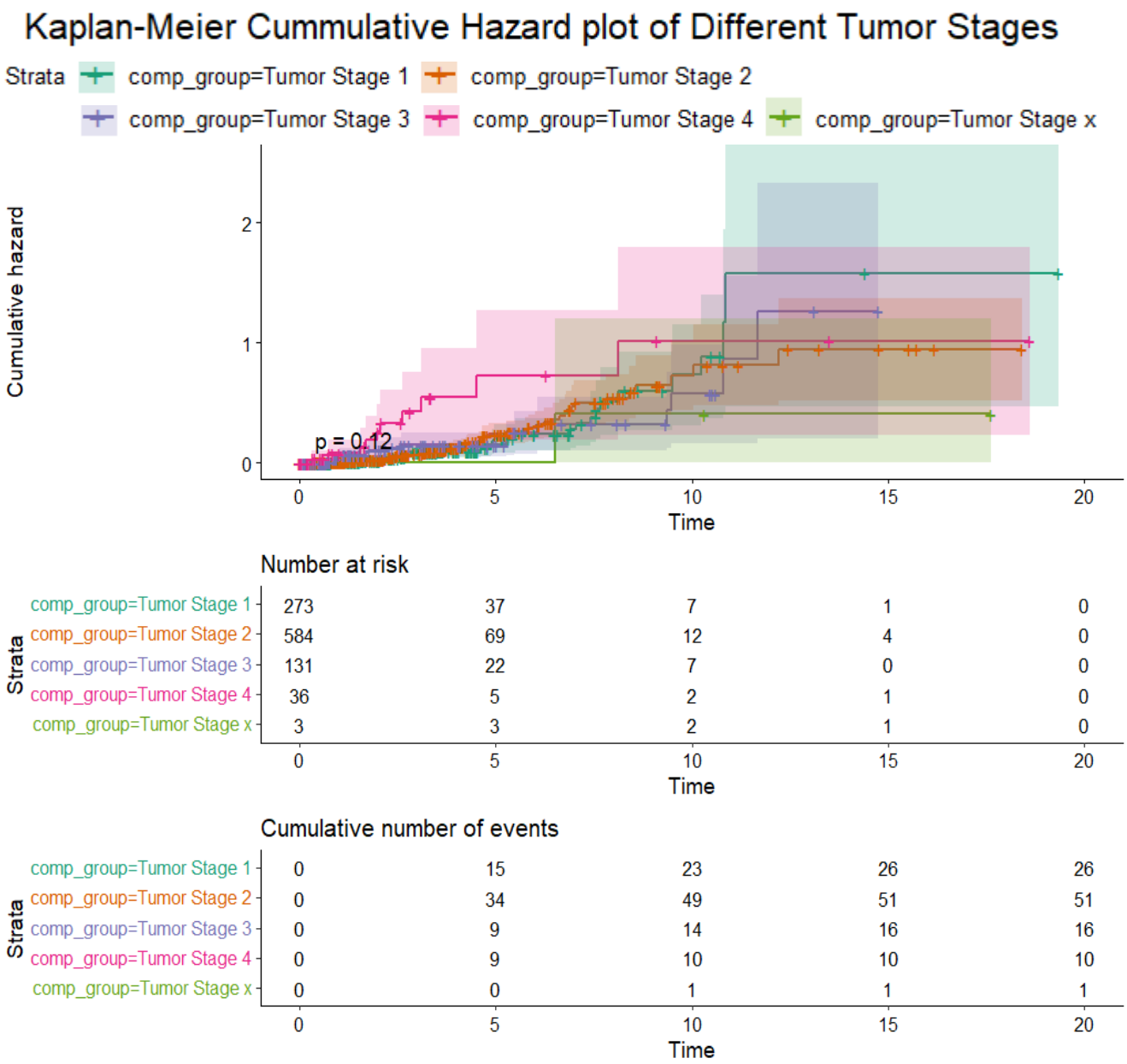
6.3.1 Non-Parametric Model fit: Kaplan-Meier

```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                              n events median 0.95LCL 0.95UCL
## comp_group=Tumor Stage 1 273     26   9.51    7.67      NA
## comp_group=Tumor Stage 2 584     51   9.48    7.82      NA
## comp_group=Tumor Stage 3 131     16  10.80    9.36      NA
## comp_group=Tumor Stage 4  36     10   4.50    2.63      NA
## comp_group=Tumor Stage x   3      1     NA    6.50      NA
```



Median Survival time for Breast Cancer patients having Stage 1 and 2 Tumors are almost equal around 9.50 Years, with Patients having Stage 3 Tumors have Median Survival time 10.8 Years.

Patients having Stage 4 Tumor have least Median Survival time i.e. 4.50 Years only.



Cumulative Hazard rate for Stage 4 tumor started quickly from early months after first diagnosis and became steady after 8 Years.

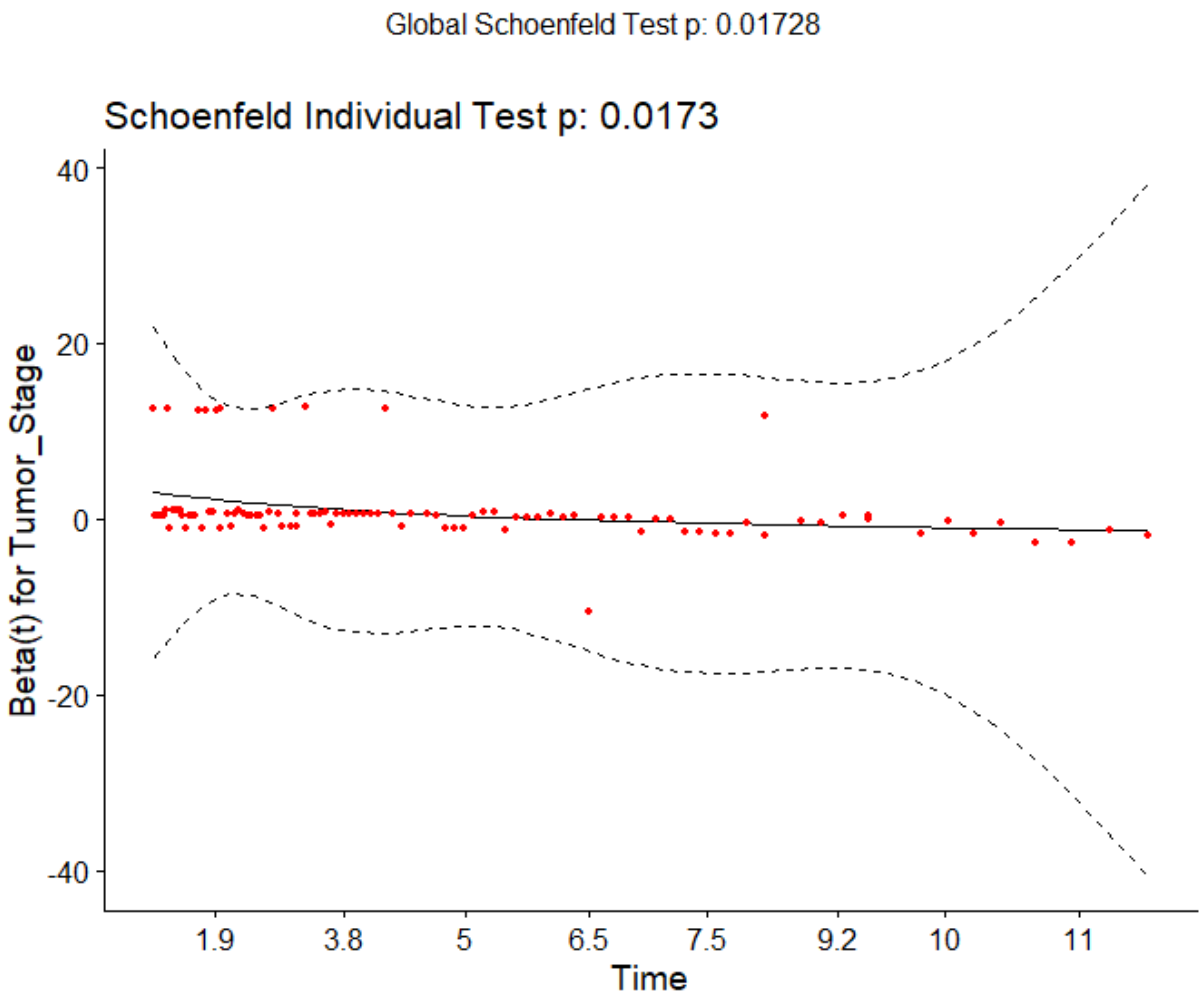
Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Tumor_Stage, data = temp)
##
##                                N Observed Expected (O-E)^2/E (O-E)^2/V
## Tumor_Stage=Tumor Stage 1 273        26    29.24  0.358630  0.502545
## Tumor_Stage=Tumor Stage 2 584        51    52.13  0.024306  0.049100
## Tumor_Stage=Tumor Stage 3 131        16    15.89  0.000726  0.000866
## Tumor_Stage=Tumor Stage 4  36        10     4.61  6.304309  6.670060
## Tumor_Stage=Tumor Stage x   3         1     2.13  0.602842  0.626232
##
##  Chisq= 7.3  on 4 degrees of freedom, p= 0.1
```

In log rank test for Difference in Survival curves we get P-value > 0.05 which indicated that We fail to reject our null hypothesis that All Survival curves from different Tumor Stages are same.

6.3.2 Semi-Parametric Model fit : Cox-Proportional Hazard Model

Proportional Hazard **Assumption** Check:



In Schoenfeld test for Proportional Hazard assumption we get global P-value as 0.0173 < 0.05.

Hence, we reject our null hypothesis of proportional hazard assumption and further we can't use Tumor Stage as a co-variate to predict Breast Cancer using Cox-PH model.

```
## Distribution | AIC      ##

## Exponential  | 866.2113 ##
## Weibull      | 827.7157 ##
## Gamma        | 823.5376 ##
## Log-Normal   | 821.5890 ##
## Log-Logistic | 818.7565 ##
```

We get lowest AIC value for **Log-Logistic Distribution**.

```
##
## Call:
## survreg(formula = Surv_obj ~ Tumor_Stage, data = temp, dist = "loglogistic")
##                              Value Std. Error     z       p
## (Intercept)                 1.5386     0.2464  6.24 4.3e-10
## Tumor_StageTumor Stage 1  0.9143     0.2751  3.32 0.00089
## Tumor_StageTumor Stage 2  0.8206     0.2616  3.14 0.00171
## Tumor_StageTumor Stage 3  0.7679     0.2967  2.59 0.00965
## Tumor_StageTumor Stage x  1.3618     0.7020  1.94 0.05239
## Log(scale)                -0.6229     0.0699 -8.91 < 2e-16
##
## Scale= 0.536
##
## Log logistic distribution
## Loglik(model)= -403.4   Loglik(intercept only)= -408.8
##  Chisq= 10.87 on 4 degrees of freedom, p= 0.028
## Number of Newton-Raphson Iterations: 8
## n= 1027
```

Took “Tumor Stage 4” as baseline survival for comparison.

Based on Chi-Square statistic and P-value < 0.05, We can say that Overall Model is statistically Significant and at least one coefficient is significantly different from 0.

As for Coefficient of Tumor Stage 1, Stage 2, and Stage 3 Patients, P-value < 0.05, We reject our null hypothesis that coefficient is 0. But for ‘Tumor Stage x’ we fail to reject our null Hypothesis.

Having ‘Tumor Stage 1’ = 1 decelerates the time to event by a factor of exp(0.9143) = 2.50 (2.50 times longer survival time compared to the baseline survival).

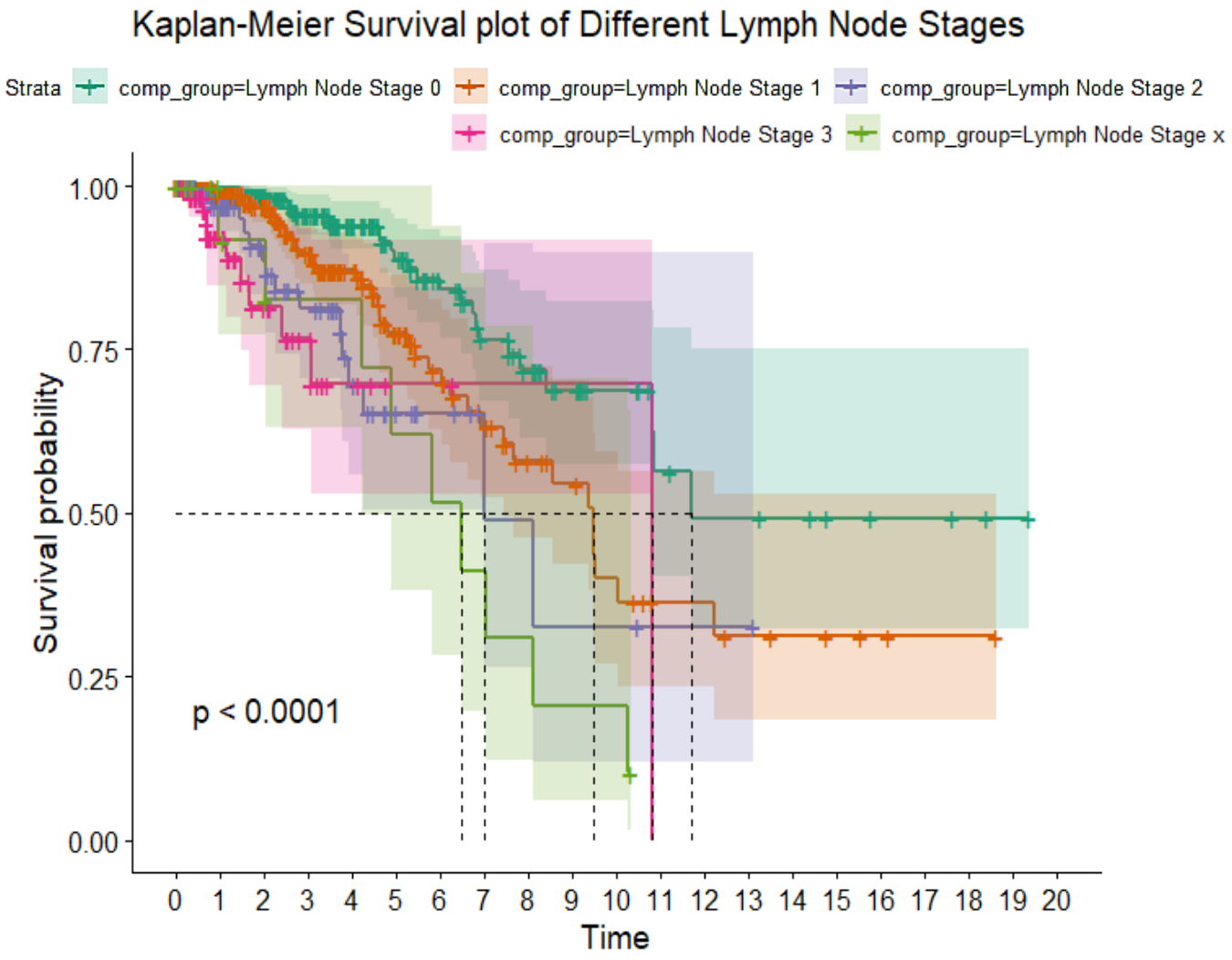
Having ‘Tumor Stage 2’ = 1 decelerates the time to event by a factor of exp(0.8206) = 2.27 (2.27 times longer survival time compared to the baseline survival).

Having ‘Tumor Stage 3’ = 1 decelerates the time to event by a factor of exp(0.7679) = 2.15 (2.15 times longer survival time compared to the baseline survival).

6.9 Lymph Node Stage

6.9.1 Non-Parametric Model fit : Kaplan-Meier

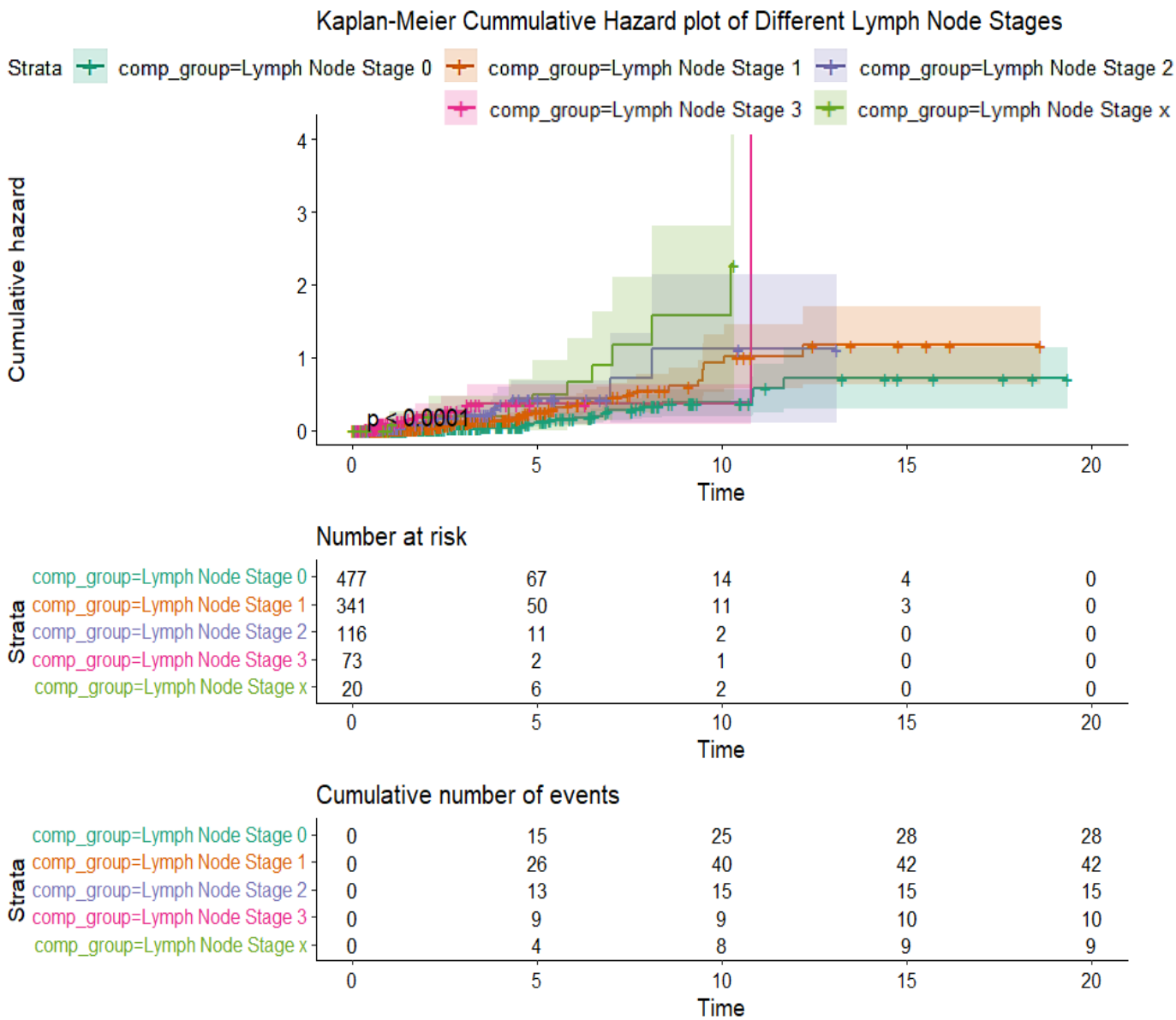
```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                                  n events median 0.95LCL 0.95UCL
## comp_group=Lymph Node Stage 0 477     28  11.69   10.81      NA
## comp_group=Lymph Node Stage 1 341     42   9.48    7.43      NA
## comp_group=Lymph Node Stage 2 116     15   6.99    6.99      NA
## comp_group=Lymph Node Stage 3  73     10  10.80      NA      NA
## comp_group=Lymph Node Stage x  20      9   6.50    4.22      NA
```



Lymph Node Stage 0 has the Highest Median Survival time that means, For this group of patients for which no cancer is found in the lymph Nodes, is likely to have median survival time of 11.69 Years.

Surprisingly Patients having Lymph node Stage 3 have 10.80 Years of Median survival time which is longer than Patients with Lymph Node Stage 1 (9.48 Years) and Lymph Node Stage 2 (6.99 Years).

For Breast Cancer Patients whose Lymph Node Stage cannot be evaluated have the least Median Suvival time i.e. 6.50 Years.



Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Lymph__node_Stage, data = temp)
##
##                                          N Observed Expected (O-E)^2/E (O-E)^2/V
## Lymph__node_Stage=Lymph Node Stage 0 477       28    49.85     9.576     18.48
## Lymph__node_Stage=Lymph Node Stage 1 341       42    38.46     0.327      0.52
## Lymph__node_Stage=Lymph Node Stage 2 116       15     8.60     4.771      5.23
## Lymph__node_Stage=Lymph Node Stage 3  73       10     3.49    12.148     12.72
## Lymph__node_Stage=Lymph Node Stage x  20        9     3.61     8.045      8.42
##
##  Chisq= 35.1  on 4 degrees of freedom, p= 4e-07
```

From High Log rank Statistic 35.1 and P-value < 0.05 we can interpret that there is significant difference in survival curves of different Lymph Node Stages and we can reject Our null hypothesis that All survival curves from different Lymph Node Stages have Same survival curve.

Pairwise-Difference in Survival Curves

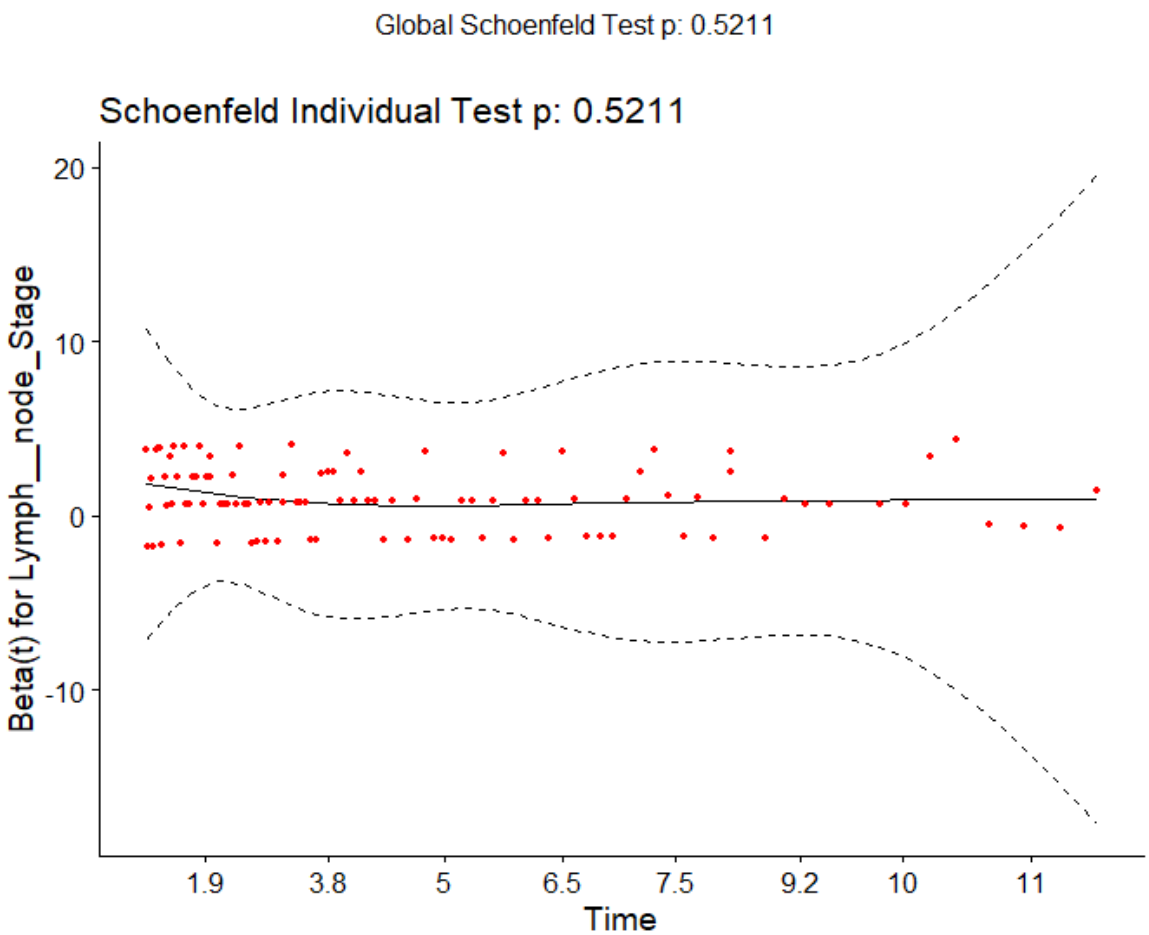
```
##              LymphNode Stage0 LymphNode Stage1 LymphNode Stage2 LymphNode Stage3
##LymphNode Stage1 *
##LymphNode Stage2 ***
##LymphNode Stage3 ****         *
##LymphNode Stagex ***          +

## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

By observing Pairwise comparison test results, We can say that Survival curve of Lymph Node Stage 0 is significantly different from Lymph Node Stages 1,2,3, and x.

Lymph NOde Stage 1 is significantly different from Stage 3 and 0.

6.9.2 Semi-Parametric Model fit : Cox-Proportional Hazard Model

Proportional Hazard **Assumption** Check:



Schoenfeld Residuals are randomly distributed around mean 0 and the test results P-value 0.5211 > 0.05.

Hence, we fail to reject our null hypothesis of proportional hazard assumption.

```
## Call:
## coxph(formula = Surv_obj ~ Lymph__node_Stage, data = temp)
##
##   n= 1027, number of events= 104
##
##                                      coef exp(coef) se(coef)     z Pr(>|z|)
## Lymph__node_StageLymph Node Stage 1 0.6682    1.9508   0.2442 2.737 0.006201
## Lymph__node_StageLymph Node Stage 2 1.1458    3.1450   0.3212 3.567 0.000360
## Lymph__node_StageLymph Node Stage 3 1.6568    5.2424   0.3736 4.435 9.22e-06
## Lymph__node_StageLymph Node Stage x 1.4935    4.4528   0.3860 3.869 0.000109
##
## Lymph__node_StageLymph Node Stage 1 **
## Lymph__node_StageLymph Node Stage 2 ***
## Lymph__node_StageLymph Node Stage 3 ***
## Lymph__node_StageLymph Node Stage x ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                     exp(coef) exp(-coef) lower .95 upper .95
## Lymph__node_StageLymph Node Stage 1     1.951     0.5126     1.209     3.148
## Lymph__node_StageLymph Node Stage 2     3.145     0.3180     1.676     5.902
## Lymph__node_StageLymph Node Stage 3     5.242     0.1908     2.521    10.903
## Lymph__node_StageLymph Node Stage x     4.453     0.2246     2.090     9.489
##
## Concordance= 0.675  (se = 0.034 )
## Likelihood ratio test= 29.61  on 4 df,   p=6e-06
## Wald test            = 31.07  on 4 df,   p=3e-06
## Score (logrank) test = 35.17  on 4 df,   p=4e-07
```

Coefficients of Lymph node stages 1,2,3, and x are statistically significant which suggests that coefficient values are different from 0.

In comparison to Lymph node Stage 0 (where either no cancer was found or Only areas of cancer smaller than 0.2mm are in the lymph nodes)-

Hazard rate for Stage 1 is 1.951 times higher. Hazard rate for Stage 2 is 3.145 times higher. Hazard rate for Stage 3 is 5.242 times higher. Hazard rate for Stage x is 4.453 times higher.

From concordance value of 0.675 and all 3 p values < 0.05 of different tests, we can conclude that overall model is statistically significant and covariate Lymph Node Stage statistically predict Hazard of Breast Cancer Patients.

6.9.3 Parametric Model fit

```
## Distribution | AIC      ##

## Exponential  | 846.0093 ##
## Weibull      | 800.7443 ##
## Gamma        | 797.9181 ##
## Log-Normal   | 801.0051 ##
## Log-Logistic | 797.7259 ##
```

We get lowest AIC value for **Log-Logistic Distribution**.

```
##
## Call:
## survreg(formula = Surv_obj ~ Lymph__node_Stage, data = temp,
##     dist = "loglogistic")
##                                      Value Std. Error     z       p
## (Intercept)                         2.6856     0.1290 20.82 < 2e-16
## Lymph__node_StageLymph Node Stage 1 -0.4164     0.1490 -2.79 0.00520
## Lymph__node_StageLymph Node Stage 2 -0.7554     0.2009 -3.76 0.00017
## Lymph__node_StageLymph Node Stage 3 -1.1078     0.2352 -4.71 2.5e-06
## Lymph__node_StageLymph Node Stage x -0.9345     0.2904 -3.22 0.00129
## Log(scale)                         -0.6579     0.0701 -9.39 < 2e-16
##
## Scale= 0.518
##
## Log logistic distribution
## Loglik(model)= -392.9   Loglik(intercept only)= -408.8
##  Chisq= 31.9 on 4 degrees of freedom, p= 2e-06
## Number of Newton-Raphson Iterations: 9
## n= 1027
```

Took “Lymph Node Stage 0” as baseline survival for comparison.

Based on Chi-Square statistic and P-value < 0.05, We can say that Overall Model is statistically Significant and at least one coefficient is significantly different from 0.

As for Coefficient of all Lymph Node Stages in comparison, P-value < 0.05, We reject our null hypothesis that coefficient is 0.

Having ‘Lymph Node Stage 1’ = 1 accelerates the time to event by a factor of exp(-0.4164) = 0.65 (0.65 times shorter survival time compared to the baseline survival).

Having ‘Lymph Node Stage 2’ = 1 accelerates the time to event by a factor of exp(-0.7554) = 0.49 (0.49 times shorter survival time compared to the baseline survival).

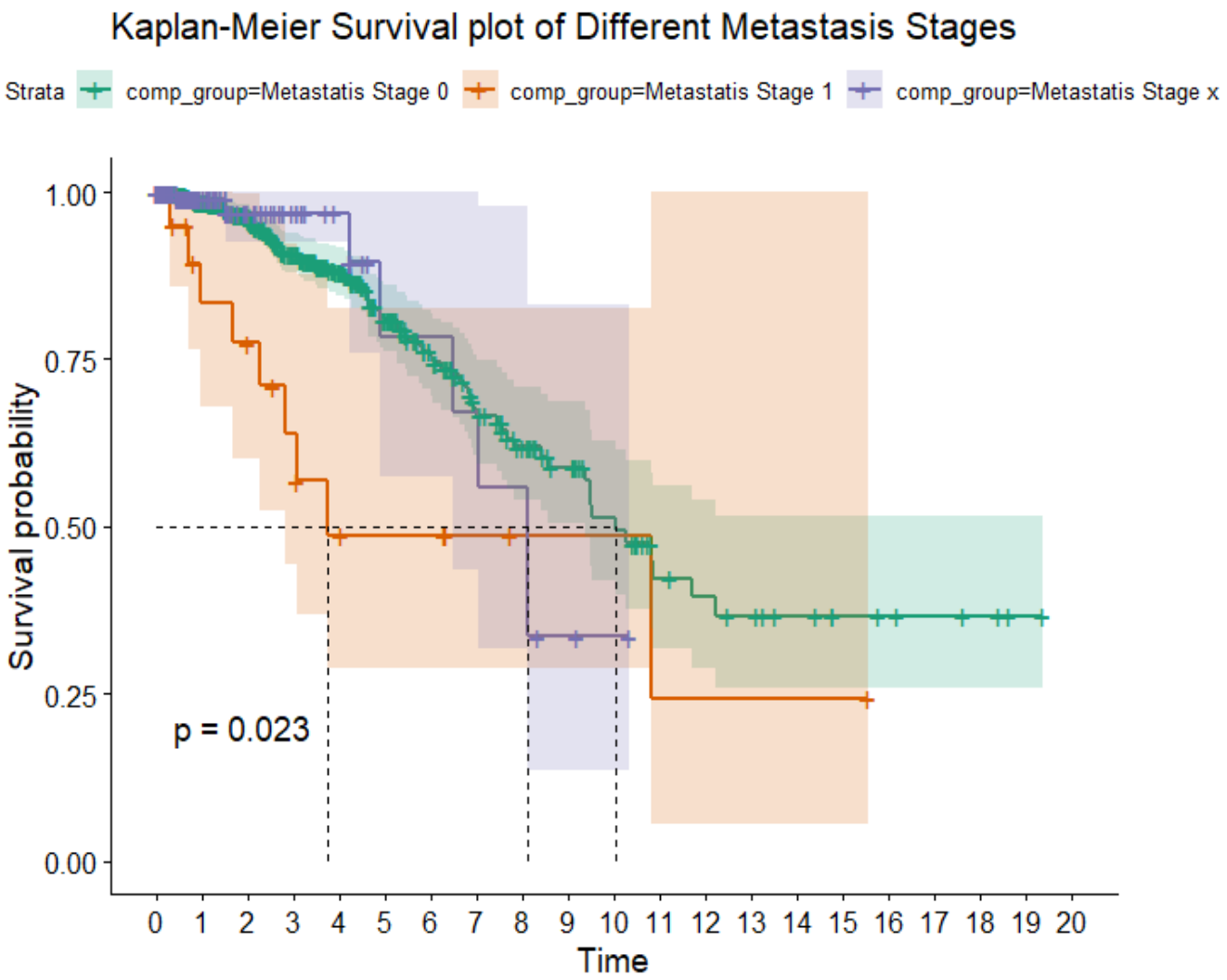
Having ‘Lymph Node Stage 3’ = 1 accelerates the time to event by a factor of exp(-1.1078) = 0.33 (0.33 times shorter survival time compared to the baseline survival).

Having ‘Lymph Node Stage x’ = 1 accelerates the time to event by a factor of exp(-0.9345) = 0.39 (0.39 times shorter survival time compared to the baseline survival).

6.10 Metastasis Stage

6.10.1 Non-Parametric Model fit : Kaplan-Meier

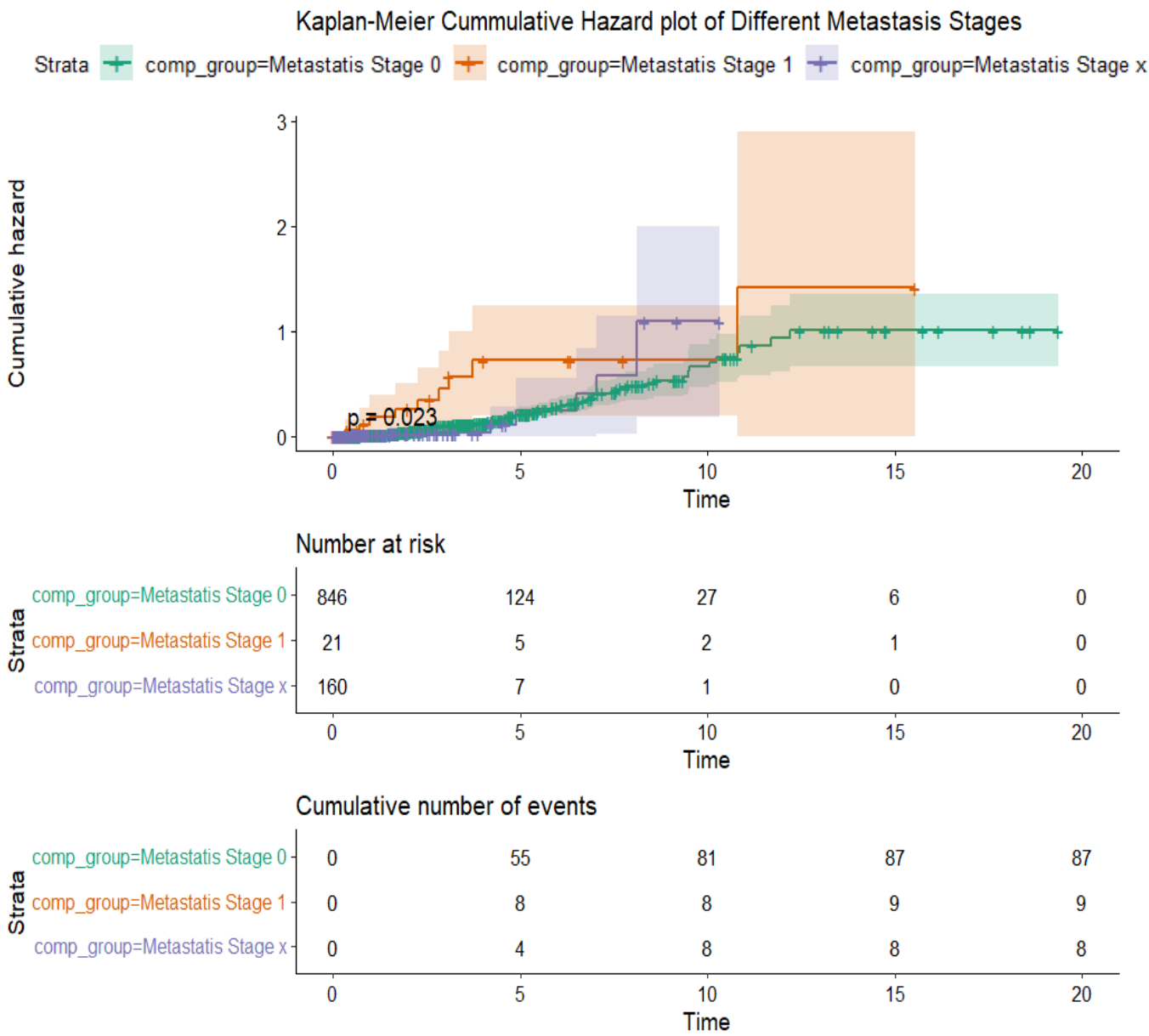
```
## Call: survfit(formula = Surv_obj ~ comp_group, data = plot_data)
##
##                                  n events median 0.95LCL 0.95UCL
## comp_group=Metastatis Stage 0 846     87  10.05    9.36      NA
## comp_group=Metastatis Stage 1  21      9   3.74    2.83      NA
## comp_group=Metastatis Stage x 160      8   8.12    6.50      NA
```



For Metastasis Stage 0 when there is no evidence of distant metastasis, Median Survival Time is 10.05 Years.

For Metastasis Stage 1 when there is evidence of metastasis to another part of the body, Median Survival time is only 3.74 Years.

For Metastasis Stage x when distant spread cannot be evaluated, Median survival time is 8.12 Years which significantly different from Median survival Time of Metastasis Stage 1.



Hazard Rate of patients having Metastasis Stage 1 Breast Cancer started increasing rapidly since first diagnosis.

Difference in Survival Curves

```
## Call:
## survdiff(formula = Surv_obj ~ Metastasis_Stage, data = temp)
##
##                                        N Observed Expected (O-E)^2/E (O-E)^2/V
## Metastasis_Stage=Metastatis Stage 0 846       87    92.44   0.32001   2.89572
## Metastasis_Stage=Metastatis Stage 1  21        9     3.78   7.20627   7.52290
## Metastasis_Stage=Metastatis Stage x 160        8     7.78   0.00619   0.00677
##
##  Chisq= 7.6  on 2 degrees of freedom, p= 0.02
```

Based on log rank statistic which follows Chi-Square distribution with 2 degrees of freedom and p-value 0.02 < 0.5, we can conclude that there is significant difference present in Metastasis Stages of Breast Cancer.

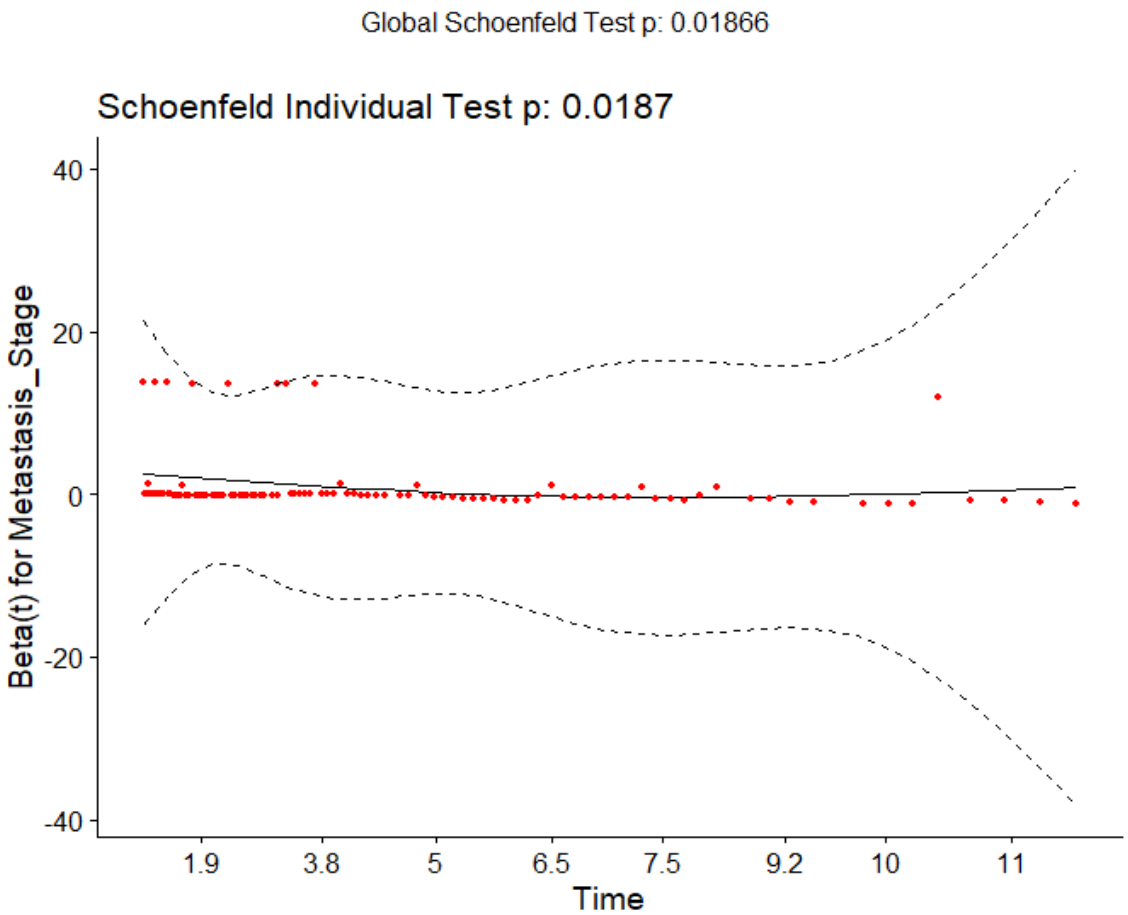
Pairwise-Difference in Survival Curves

```
##                     Metastatis Stage 0 Metastatis Stage 1
## Metastatis Stage 1 *
## Metastatis Stage x                    *
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

There is significant difference between all the Survival curves of different Metastasis Stage pairs.

6.10.2 Semi-Parametric Model fit : Cox-Proportional Hazard Model

Proportional Hazard **Assumption** Check:



Based on Schoenfeld Test and P-value 0.0187 < 0.05, We have to reject our null hypothesis of proportional hazard assumption. Hence, we cannot use Metastasis Stage as a co-variate to predict Hazard of Breast Cancer Patients.

6.10.3 Parametric Model fit

```
## Distribution | AIC      ##

## Exponential  | 861.7261 ##
## Weibull      | 823.5559 ##
## Gamma        | 819.9043 ##
## Log-Normal   | 819.2151 ##
## Log-Logistic | 816.4410 ##
```

We get lowest AIC value for **Log-Logistic Distribution**.

```
##
## Call:
## survreg(formula = Surv_obj ~ Metastasis_Stage, data = temp, dist = "loglogistic
")
##                                      Value Std. Error     z      p
## (Intercept)                         2.3897     0.0878 27.22 <2e-16
## Metastasis_StageMetastatis Stage 1 -0.8954     0.2806 -3.19 0.0014
## Metastasis_StageMetastatis Stage x -0.0060     0.2292 -0.03 0.9791
## Log(scale)                         -0.6135     0.0699 -8.78 <2e-16
##
## Scale= 0.541
##
## Log logistic distribution
## Loglik(model)= -404.2   Loglik(intercept only)= -408.8
##  Chisq= 9.18 on 2 degrees of freedom, p= 0.01
## Number of Newton-Raphson Iterations: 7
## n= 1027
```

Took “Metastasis Stage 0” as baseline survival for comparison.

Based on Chi-Square statistic and P-value < 0.05, We can say that Overall Model is statistically Significant and at least one coefficient is significantly different from 0.

As for Coefficient of Metastasis Stage 1 Patients, P-value < 0.05, We reject our null hypothesis that coefficient is 0. But for Metastasis Stage x we fail to reject null hypothesis.

Having ‘Metastasis Stage 1’ = 1 accelerates the time to event by a factor of exp(-0.8954) = 0.41 (0.41 times shorter survival time compared to the baseline survival).

6.11 Multivariate Cox-Proportional Model

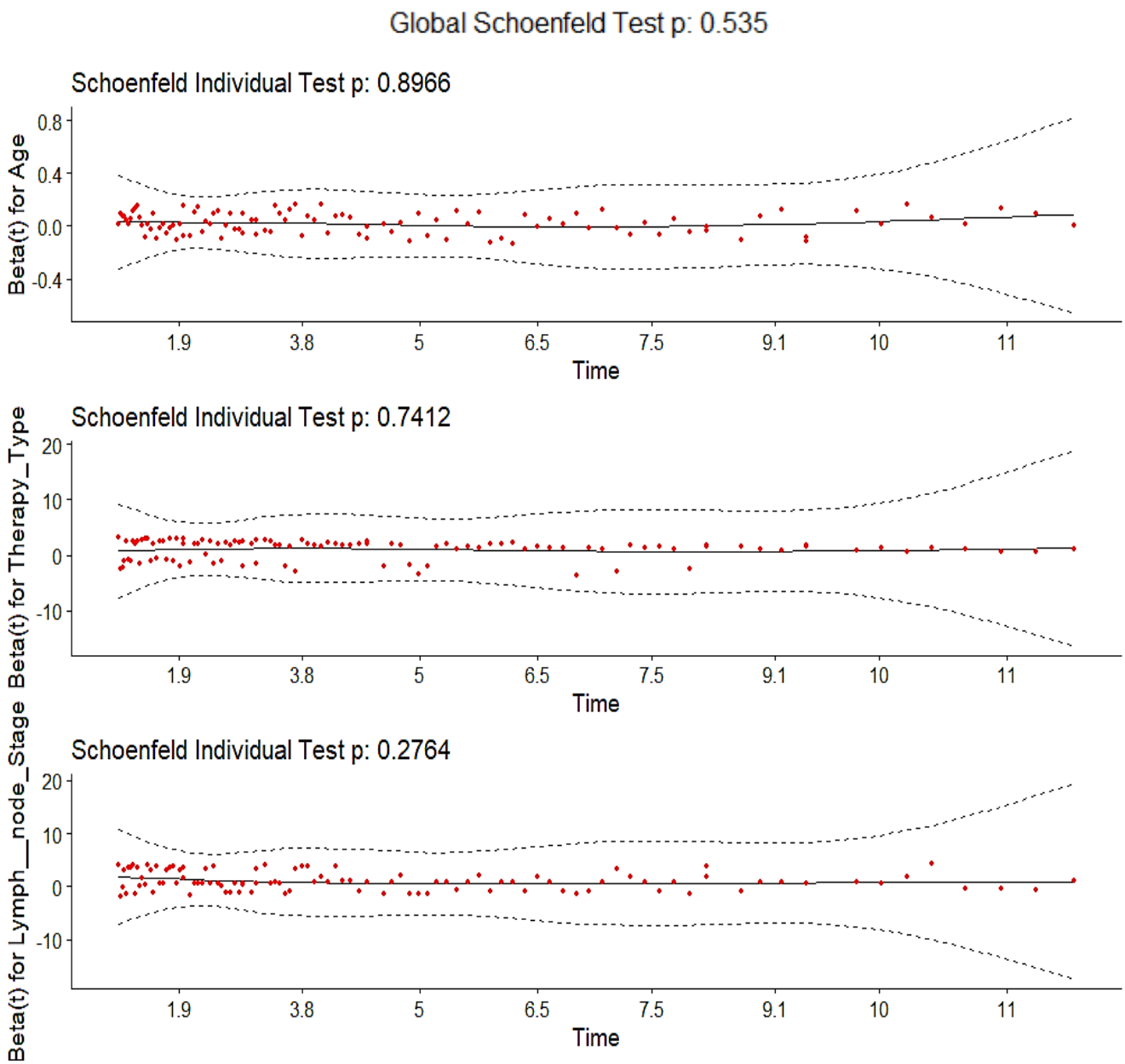
Based on Univariate analysis results these covariates are found to be contributing in predicting Survival of Breast Cancer Patients using Cox-Proportional Model.

Age Therapy Type Lymph Node

Semi-Parametric Model fit : Cox-Proportional Hazard Model

```
##                      chisq df    p
## Age                 0.0169  1 0.90
## Therapy_Type        1.2493  3 0.74
## Lymph__node_Stage 5.1083  4 0.28
## GLOBAL              7.0151  8 0.54
```

Proportional Hazard **Assumption** Check:



From Individual Schoenfeld Test, Schoenfeld residuals are randomly distributed around mean 0 and Global P-value 0.535 > 0.05 which indicates that proportional hazard assumption is met for all these covariates and we can use these variables in our final Cox-PH model to predict Hazard of Breast Cancer Patients.

```
summary(fit_coxph)

## Call:
## coxph(formula = Surv_obj ~ Age + Therapy_Type + Lymph__node_Stage,
##     data = temp)
##
##   n= 1018, number of events= 104
##
#                                        coef exp(coef) se(coef)     z Pr(>|z|)
# Age                                 0.02061  1.02083 0.00758  2.719  0.00654 **
# Therapy_Typehormone therapy        -0.29539  0.74424 0.45733 -0.646  0.51835
# Therapy_TypeNo Information          1.35314  3.86954 0.28645  4.724 2.31e-06 ***
# Therapy_TypeOther                   0.45496  1.57611 1.03003  0.442  0.65871
# Lymph__node_StageLymph Node Stage 1 0.54868  1.73096 0.24618  2.229  0.02583 *
# Lymph__node_StageLymph Node Stage 2 1.46699  4.33616 0.32742  4.480 7.45e-06 ***
# Lymph__node_StageLymph Node Stage 3 1.64808  5.19698 0.37430  4.403 1.07e-05 ***
# Lymph__node_StageLymph Node Stage x 0.85684  2.35569 0.39719  2.157  0.03098 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                      exp(coef) exp(-coef) lower .95 upper .95
## Age                                     1.0208     0.9796    1.0058     1.036
## Therapy_Typehormone therapy             0.7442     1.3436    0.3037     1.824
## Therapy_TypeNo Information              3.8695     0.2584    2.2072     6.784
## Therapy_TypeOther                       1.5761     0.6345    0.2093    11.867
## Lymph__node_StageLymph Node Stage 1     1.7310     0.5777    1.0684     2.804
## Lymph__node_StageLymph Node Stage 2     4.3362     0.2306    2.2825     8.238
## Lymph__node_StageLymph Node Stage 3     5.1970     0.1924    2.4955    10.823
## Lymph__node_StageLymph Node Stage x     2.3557     0.4245    1.0815     5.131
##
## Concordance= 0.774  (se = 0.028 )
## Likelihood ratio test= 80.39  on 8 df,   p=4e-14
## Wald test            = 74.99  on 8 df,   p=5e-13
## Score (logrank) test = 85.73  on 8 df,   p=3e-15
```

Concordance value indicated this Cox-PH model has predicting ability of 77 %.

Final Cox-Proportional Model Equation for Breast Cancer Patients –

$$H(t|Age, No\ Information, Lymph\ Node\ Stage\ 1, Lymph\ Node\ Stage\ 2, Lymph\ Node\ Stage\ 3, Lymph\ Node\ Stage\ X) =$$

$$H0(t)exp(0.02 * Age + 1.35 * No\ Information + 0.54 * Lymph\ Node\ Stage\ 1 + 1.47 * Lymph\ Node\ Stage\ 2 + 1.65 * Lymph\ Node\ Stage\ 4 + 0.85 * Lymph\ Node\ Stage\ X)$$

7. Results

I. **No Covariate:** Median Survival Time is 9.51 Years.

II. Age

a. *Kaplan Meier Model fit*: Median Survival time for the Middle age group is highest, 12.21 Years with Young age having the least Median survival time, 8.39 Years, and Middle age having 9.36 Years.
A significant difference between the Middle Age and Old Age groups is found.

b. *Cox-PH Model fit:* For every 1-year increase in age the risk of death will increase by 2.7%.

c. *Parametric Model fit:* Having 'Old Age' = 1 accelerates the time to event by a factor of 0.650 (0.650 times shorter survival time compared to the baseline survival Median Age).

III. Race

a. *Kaplan Meier Model fit:* Median survival time for Black or African American and White group is same around 9.50 Years. No Difference in Survival Curves.

b. *Cox-PH Model fit*: Not a Good fit. (Statistically Insignificant)

c. *Parametric Model fit:* Not a Good fit. (Statistically Insignificant)

IV. Ethnicity

a. *Kaplan Meier Model fit:* Median survival time for Not Hispanic or Latino group is 9.51 Years while we KM-Model failed to calculated Median Survival time for Hispanic or Latino group.

b. *Cox-PH Model fit:* Not a Good fit. (Coefficient value is Infinite)

c. *Parametric Model fit:* Not a Good fit.

V. Therapy Type

a. *Kaplan Meier Model fit:* Median Survival time is 6.9 Years for No Therapy type. For Hormone Therapy, Chemotherapy and Other therapies, KM-Model failed to compute Median Survival times.
Significant difference is found between No therapy and Chemotherapy, Hormone Therapy.

b. *Cox-PH Model fit:* Cumulative Hazard rate for no information group is 4 times larger than chemotherapy.

c. *Parametric Model fit:* Having 'Chemotherapy' = 1 accelerates the time to event by a factor of 2.54 (2.54 times longer survival time compared to the baseline survival No Therapy).
Having 'Hormone Therapy' = 1 accelerates the time to event by a factor of 3.06 (3.06 times longer survival time compared to the baseline survival No Therapy).

VI. Cancer Stage

a. *Kaplan Meier Model fit:* No Significant difference in Survival of Stage 1 and 2 Cancer. Other Cancer Stages have significant difference in their Survival. Stage with Stage 3,4, and X having median Survival Time 9.36, 3.74, and 6.50 respectively.

b. *Cox-PH Model fit:* Proportional Hazard assumption not met.

c. *Parametric Model fit:* Having 'Stage 3' = 1 accelerates the time to event by a factor of 0.50 (0.50 times shorter survival time compared to the baseline survival Stage 1).
Having 'Stage 4' = 1 accelerates the time to event by a factor of 0.24 (0.24 times shorter survival time compared to the baseline survival Stage 1).
Having 'Stage x' = 1 (When Cancer Stage cannot be assessed) accelerates the time to event by a factor of 0.42 (0.42 times shorter survival time compared to the baseline survival Stage 1).

VII. Tumor Stage (T)

a. *Kaplan Meier Model fit:* No significant difference in Survival of Tumor Stage 1 and 2.
Patients having Stage 3 Tumors have Median Survival time 10.8 Years.
Patients having Stage 4 Tumor have least Median Survival time- 4.50 Years only.

b. *Cox-PH Model fit:* Not a Good fit. (Statistically Insignificant)

c. *Parametric Model fit*: Having 'Tumor Stage 1' = 1 decelerates the time to event by a factor of 2.50 (2.50 times longer survival time compared to the baseline survival Tumor Stage 4).
Having 'Tumor Stage 2' = 1 decelerates the time to event by a factor of 2.27 (2.27 times longer survival time compared to the baseline survival Tumor Stage 4).
Having 'Tumor Stage 3' = 1 decelerates the time to event by a factor of 2.15 (2.15 times longer survival time compared to the baseline survival Tumor Stage 4).

VIII. Lymph Node Stage (N)

a. *Kaplan Meier Model fit:* No significant difference in Survival of Lymph Node Stage 2 and 3.
Survival for Lymph Node Stage 0 is significantly different from all other Stages.
Median Survival time for Stage 0,1,2, and X is 11.69, 9.48, 6.99, 10.80, 6.50 Years respectively.

b. *Cox-PH Model fit:* In comparison to Lymph Node Stage 0, Hazard rate for Stage 1 is 1.951 times higher. Hazard rate for Stage 2 is 3.145 times higher. Hazard rate for Stage 3 is 5.242 times higher. Hazard rate for Stage x is 4.453 times higher.

c. *Parametric Model fit*: Having 'Lymph Node Stage 1' = 1 accelerates the time to event by a factor of 0.65 (0.65 times shorter survival time compared to the baseline survival).
Having 'Lymph Node Stage 2' = 1 accelerates the time to event by a factor of 0.49 (0.49 times shorter survival time compared to the baseline survival Lymph Node Stage 0).
Having 'Lymph Node Stage 3' = 1 accelerates the time to event by a factor of 0.33 (0.33 times shorter survival time compared to the baseline survival Lymph Node Stage 0).
Having 'Lymph Node Stage x' = 1 accelerates the time to event by a factor of 0.39 (0.39 times shorter survival time compared to the baseline survival Lymph Node Stage 0).

IX. Metastasis Stage (M)

a. *Kaplan Meier Model fit:* Significant Difference is found between Metastasis Stage 0 and 1.
For Metastasis Stage 0 and X, Median Survival Time is 10.05 and 8.12 Years.
For Metastasis Stage 1, Median Survival Time is 3.74 Years.

b. *Cox-PH Model fit:* Not a good fit. (Statistically Insignificant)

c. *Parametric Model fit*: Having 'Metastasis Stage 1' = 1 accelerates the time to event by a factor of 0.41 (0.41 times shorter survival time compared to the baseline survival Metastasis Stage 0).

X. Multivariate Cox-Proportional Model

The baseline Cox-PH model corresponds to -

Age = ‘0’,
Therapy Type = ‘Chemotherapy’,
Lymph Node Stage = ‘Lymph Node Stage 0’.

By every 1-year increase in ‘Age’ accelerates the time to event by a factor of 1.027.

Compared to patients who took 'Chemotherapy, Patients who didn’t take any Breast Cancer Therapy (‘No Therapy’) accelerates the time to event by a factor of 3.86.

Compared to patients having Lymph Node Stage 0, Patients having ‘Lymph Node Stage 1’ accelerates the time to event by a factor of `1.73`.

Compared to patients having Lymph Node Stage 0, Patients having ‘Lymph Node Stage 2’ accelerates the time to event by a factor of 4.34.

Compared to patients having Lymph Node Stage 0, Patients having ‘Lymph Node Stage 3’ accelerates the time to event by a factor of 5.19.

Compared to patients having Lymph Node Stage 0, Patients having ‘Lymph Node Stage X’ accelerates the time to event by a factor of 2.36.

8. Appendix: R Code

```
##################################################
##################################################
## TCGA-BRCA Clinical Life-Time Data Analysis ##
##################################################
##################################################

####################
# Libraries needed #
####################

# if (!requireNamespace("BiocManager", quietly = TRUE))
#   install.packages("BiocManager")
# BiocManager::install()

library(RTCGA)
library(RTCGA.clinical)
library(RTCGA.mRNA)
library(tidyverse)
library(Hmisc)
library(ggplot2)
library(pivottabler)
library(survminer)
library(survival)
library(writexl)
library(ggplot2)
library(SurvRegCensCov)
library(flexsurv)
################
# Data Loading #
################

#Loading Data for comparison of cancers prone to women
cancer_clin_data = survivalTCGA(UCEC.clinical, BRCA.clinical, OV.clinical, CESC.cl
inical, extract.cols = c("admin.disease_code","patient.drugs.drug.therapy_types.th
erapy_type"))
dim(cancer_clin_data)

#Out of 4 cancer types women's are more prone to..
sort(table(cancer_clin_data$admin.disease_code),decreasing = T)

#Loading Data for Breast Cancer
BRCA_data = survivalTCGA(BRCA.clinical,
                         extract.cols = c("patient.gender",
                                          "patient.race",
                                          "patient.ethnicity",
                                          "patient.days_to_birth",
                                          "patient.drugs.drug.therapy_types.therap
y_type",
                                          "patient.stage_event.pathologic_stage",
                                          "patient.stage_event.tnm_categories.path
```

```
ologic_categories.pathologic_t",
                                          "patient.stage_event.tnm_categories.path
ologic_categories.pathologic_n",
                                          "patient.stage_event.tnm_categories.path
ologic_categories.pathologic_m"))

dim(BRCA_data)
# write_xlsx(cancer_clin_data,"D:/Project/TCGA-BRCA Clinical LIfe-Time Data Analys
is/cancer_clin_data.xlsx")
# write_xlsx(BRCA_data,"D:/Project/TCGA-BRCA Clinical LIfe-Time Data Analysis/BRCA
_data.xlsx")

#######################
# Data Pre-processing #
#######################

#Firstly we will rename the long variables names to short and meaningful names
cancer_clin_data = cancer_clin_data %>% rename(Disease_Code = admin.disease_code,
                                               Patient_code = bcr_patient_barcode,
                                               Survival_Time = times,
                                               Vital_Status = patient.vital_status
,
                                               Therapy_Type = patient.drugs.drug.t
herapy_types.therapy_type
)

BRCA_data = BRCA_data %>% rename(Patient_code = bcr_patient_barcode,
                                 Survival_Time = times,
                                 Vital_Status = patient.vital_status,
                                 Gender = patient.gender,
                                 Race = patient.race,
                                 Ethnicity = patient.ethnicity,
                                 Age = patient.days_to_birth,
                                 Therapy_Type = patient.drugs.drug.therapy_types.t
herapy_type,
                                 Cancer_Stage = patient.stage_event.pathologic_sta
ge,
                                 Tumor_Stage = patient.stage_event.tnm_categories.
pathologic_categories.pathologic_t,
                                 Lymph__node_Stage = patient.stage_event.tnm_categ
ories.pathologic_categories.pathologic_n,
                                 Metastasis_Stage = patient.stage_event.tnm_catego
ries.pathologic_categories.pathologic_m
)

head(cancer_clin_data)
str(cancer_clin_data)
# describe(BRCA_data)

head(BRCA_data)
str(BRCA_data)
```

```
#We need to Convert Survival Times from number of days to Years.
cancer_clin_data$Survival_Time = round((cancer_clin_data$Survival_Time/365),2)
BRCA_data$Survival_Time = round((BRCA_data$Survival_Time/365),2)

#Need to make Age variable more interpretable.
BRCA_data$Age = abs(round(as.numeric(BRCA_data$Age)/365,2))

BRCA_data = mutate(BRCA_data,
                   Age_Category = cut(Age, breaks = c(0,40,60,Inf),labels = c("You
ng Age","Middle Age","Old Age")))

#As the data is right censored checking and Removing those patients who has negati
ve survival time
cancer_clin_data = cancer_clin_data %>% filter(Survival_Time>0)
BRCA_data = BRCA_data %>% filter(Survival_Time>0)

#As this study is about Cancer prone to females then we'll be removing 12 male pat
ients
BRCA_data = BRCA_data %>% filter(Gender == "female")

#Checking for duplicate patient data
length(unique(cancer_clin_data$Patient_code))
length(unique(BRCA_data$Patient_code))

#Race
table(BRCA_data$Race)
BRCA_data = BRCA_data[-which(BRCA_data$Race == "american indian or alaska native")
,]

#Therapy Type
table(BRCA_data$Therapy_Type)
BRCA_data$Therapy_Type[is.na(BRCA_data$Therapy_Type)] = "No Information"
BRCA_data$Therapy_Type = fct_lump(BRCA_data$Therapy_Type,n = 3)
table(cancer_clin_data$Therapy_Type)
cancer_clin_data$Therapy_Type[is.na(cancer_clin_data$Therapy_Type)] = "No Informat
ion"
cancer_clin_data$Therapy_Type = fct_lump(cancer_clin_data$Therapy_Type,n = 3)

#We need to group, subgroups of Cancer Stage
BRCA_data$Cancer_Stage = str_replace_all(BRCA_data$Cancer_Stage,"stage iii[a-c]|st
age iii$","Stage 3")
BRCA_data$Cancer_Stage = str_replace_all(BRCA_data$Cancer_Stage,"stage ii[a-b]|sta
ge ii$","Stage 2")
BRCA_data$Cancer_Stage = str_replace_all(BRCA_data$Cancer_Stage,"stage i[a-b]|stag
e i$","Stage 1")
BRCA_data$Cancer_Stage = str_replace_all(BRCA_data$Cancer_Stage,"stage iv$","Stage
4")
BRCA_data$Cancer_Stage = str_replace_all(BRCA_data$Cancer_Stage,"stage x$","Stage
x")

#We need to group, subgroups of Tumor Stage
```

```
BRCA_data$Tumor_Stage = str_replace_all(BRCA_data$Tumor_Stage,"t4[a-d]|t4$","Tumor
Stage 4")
BRCA_data$Tumor_Stage = str_replace_all(BRCA_data$Tumor_Stage,"t3[a-d]|t3$","Tumor
Stage 3")
BRCA_data$Tumor_Stage = str_replace_all(BRCA_data$Tumor_Stage,"t2[a-d]|t2$","Tumor
Stage 2")
BRCA_data$Tumor_Stage = str_replace_all(BRCA_data$Tumor_Stage,"t1[a-d]|t1$","Tumor
Stage 1")
BRCA_data$Tumor_Stage = str_replace_all(BRCA_data$Tumor_Stage,"tx$","Tumor Stage x
")

# We need to group, subgroups of Lymph Stage
BRCA_data$Lymph__node_Stage = str_replace_all(BRCA_data$Lymph__node_Stage,"^n3[a-d
]|n3","Lymph Node Stage 3")
BRCA_data$Lymph__node_Stage = str_replace_all(BRCA_data$Lymph__node_Stage,"^n2[a-d
]|n2","Lymph Node Stage 2")
BRCA_data$Lymph__node_Stage = str_replace_all(BRCA_data$Lymph__node_Stage,"^n1[a-d
]|n1mi|n1","Lymph Node Stage 1")
BRCA_data$Lymph__node_Stage = str_replace_all(BRCA_data$Lymph__node_Stage,"n0\\s+\
\([a-z]*[\\+|-]\\)|n0","Lymph Node Stage 0")
BRCA_data$Lymph__node_Stage = str_replace_all(BRCA_data$Lymph__node_Stage,"^nx","L
ymph Node Stage x")

#We need to group, subgroups of Metastatis Stage
BRCA_data$Metastasis_Stage = str_replace_all(BRCA_data$Metastasis_Stage,"cm0\\s+\\
([a-z]*[\\+|-]\\)|m0","Metastatis Stage 0")
BRCA_data$Metastasis_Stage = str_replace_all(BRCA_data$Metastasis_Stage,"m1","Meta
statis Stage 1")
BRCA_data$Metastasis_Stage = str_replace_all(BRCA_data$Metastasis_Stage,"^mx","Met
astatis Stage x")

#Proper Data types
str(BRCA_data)
BRCA_data[,c(4:6,8:12)] = lapply(BRCA_data[,c(4:6,8:12)],factor)

# write_xlsx(cancer_clin_data,"D:/Project/TCGA-BRCA Clinical LIfe-Time Data Analys
is/P_PreProcessed_cancer_clin_data.xlsx")
# write_xlsx(BRCA_data,"D:/Project/TCGA-BRCA Clinical LIfe-Time Data Analysis/P_Pr
eProcessed_BRCA_data.xlsx")

head(BRCA_data)

##############################
##Exploratory Data Analysis##
##############################

### Distribution of Vital Status

ggplot(data = cancer_clin_data,aes(x = factor(Vital_Status)))+
  geom_bar(aes(fill = factor(Vital_Status))) +
  labs(title = "Vital Status Distribution | Distribution of Cancers - Prone to Fem
```

```
ales") +
  scale_fill_discrete(name = "Vital Status", labels = c("0 - Censored (Alive / Los
t to follow-up / Accidental Death)",
                                                        "1 - Event (Death)"))

### Distribution of type of Cancers only Prone to females

prop.table(table(cancer_clin_data$Disease_Code))*100

cancer_clin_data %>%
  arrange(Disease_Code) %>%
  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Disease_Code,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  labs(x = "Survival Time", y = "Patients") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
th)")) +
  scale_color_discrete(name = " ", labels = c("BRCA - Breast Cancer",
                                              "CESC - Cervical Cancer",
                                              "OV - Ovarian Cancer",
                                              "UCEC - Uterine Cancer"))


We can see out these four cancer types, 42% cases were alone having Breast Cancer
and*
  Cervical Cancer is the least prone out of these.*

### EDA for Breast Cancer

### Distribution of Vital Status | Breast Cancer

paste(round(prop.table(table(BRCA_data$Vital_Status))*100),c("% Patients have cens
ored survival time",
                                                              "% Patients have conf
irmed survival time"))

ggplot(data = BRCA_data,aes(x = factor(Vital_Status)))+
  geom_bar(aes(fill = factor(Vital_Status))) +
  labs(title = "Vital Status Distribution | Breast Cancer") +
  scale_fill_discrete(name = "Vital Status", labels = c("0 - Censored (Alive / Los
t to follow-up / Accidental Death)",
                                                        "1 - Event (Death)"))

### Survival Plot with respect to Age Category | Breast Cancer
```

```
qhpvt(BRCA_data,"Vital_Status","Age_Category","n()")

prop.table(table(BRCA_data$Vital_Status,BRCA_data$Age_Category))*100

BRCA_data %>%
  arrange(Age_Category) %>%
  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Age_Category,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  ggtitle("Survival Plot with respect to Age Category | Breast Cancer") +
  labs(x="Survival Time after first diagnosis (Years)", y="Patients") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
th)")) +
  scale_color_discrete(name = "Age Catgory", labels = c("Young Age (0-40)",
                                                        "Middle Age (40-60)",
                                                        "Old Age (60+)",
                                                        "No Information"))

### Survival Plot with respect to Race | Breast Cancer

qhpvt(BRCA_data,"Vital_Status","Race","n()")

BRCA_data %>%
  arrange(Race) %>%
  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Race,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  ggtitle("Survival Plot with respect to Race | Breast Cancer") +
  labs(x="Survival Time after first diagnosis (Years)", y="Patients") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
th)")) +
  scale_color_discrete(name = "Race", labels = c("Black or African american",
                                                 "White",
                                                 "Other",
                                                 "No Information"))

### Survival Plot with respect to Ethnicity | Breast Cancer
```

```
qhpvt(BRCA_data,"Vital_Status","Ethnicity","n()")

BRCA_data %>%
  arrange(Ethnicity) %>%
  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Ethnicity,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  ggtitle("Survival Plot with respect to Ethnicity | Breast Cancer") +
  labs(x="Survival Time after first diagnosis (Years)", y="Patients") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
th)")) +
  scale_color_discrete(name = "Ethnicity", labels = c("Hispanic or Latino",
                                                      "Not Hispanic or Latino",
                                                      "No Information Available"))

### Survival Plot with respect to Therapy Type | Breast Cancer

qhpvt(BRCA_data,"Vital_Status","Therapy_Type","n()")

BRCA_data %>%
  arrange(Therapy_Type) %>%
  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Therapy_Type,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  ggtitle("Survival Plot with respect to Therapy Type | Breast Cancer") +
  labs(x="Survival Time after first diagnosis (Years)", y="Patients") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
th)")) +
  scale_color_discrete(name = "Therapy Type", labels = c("chemotherapy",
                                                         "hormone therapy",
                                                         "Other",
                                                         "No Information Available
"))

### Survival Plot with respect to Cancer Stage | Breast Cancer

qhpvt(BRCA_data,"Vital_Status","Cancer_Stage","n()")
```

```
# prop.table(table(BRCA_data$Vital_Status,BRCA_data$Cancer_Stage))

BRCA_data %>%
  arrange(Cancer_Stage) %>%

  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Cancer_Stage,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  ggtitle("Survival Plot with respect to Cancer Stage | Breast Cancer") +
  labs(x="Survival Time after first diagnosis (Years)", y="Patients") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
th)")) +
  scale_color_discrete(name = "Cancer Stage", labels = c("Stage 1",
                                                         "Stage 2",
                                                         "Stage 3",
                                                         "Stage 4",
                                                         "Stage x",
                                                         "No Information Available
"))

#Therapy Type ~ Cancer Stage

qhpvt(BRCA_data,"Therapy_Type","Cancer_Stage","n()")

#Survival Plot with respect to Tumor Stage | Breast Cancer

qhpvt(BRCA_data,"Vital_Status","Tumor_Stage","n()")

BRCA_data %>%
  arrange(Tumor_Stage) %>%
  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Tumor_Stage,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  ggtitle("Survival Plot with respect to Tumor Stage | Breast Cancer") +
  labs(x="Survival Time after first diagnosis (Years)", y="Patients") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
```

```
th)"))

#Cancer Stage ~ Tumor Stage

qhpvt(BRCA_data,"Cancer_Stage","Tumor_Stage","n()")

#Survival Plot with respect to Lymph Node Stage | Breast Cancer

qhpvt(BRCA_data,"Vital_Status","Lymph__node_Stage","n()")

BRCA_data %>%
  arrange(Lymph__node_Stage) %>%
  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Lymph__node_Stage,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  ggtitle("Survival Plot with respect to Lymph node Stage | Breast Cancer") +
  labs(x="Survival Time after first diagnosis (Years)", y="Patients") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
th)"))

#Cancer Stage ~ Lymph Node Stage

qhpvt(BRCA_data,"Cancer_Stage","Lymph__node_Stage","n()")

### Survival Plot with respect to Metastasis Stage | Breast Cancer

qhpvt(BRCA_data,"Vital_Status","Metastasis_Stage","n()")

BRCA_data %>%
  arrange(Metastasis_Stage) %>%
  mutate(index=1:n()) %>%
  ggplot(
    aes(xend = 0,
        y = index,
        x = Survival_Time,
        yend = index,
        colour = Metastasis_Stage,
        shape = factor(Vital_Status))) +
  geom_segment(size = .02) +
  geom_point() +
  labs(x="Survival Time after first diagnosis (Years)",
       y="Patients",
       title = "Survival Plot with respect to Metastasis Stage | Breast Cancer") +
  scale_shape_discrete(name = "Survival Status", labels = c("Censored","Event (Dea
```

```
th)"))

##### **Cancer Stage ~ Metastasis Stage**

qhpvt(BRCA_data,"Cancer_Stage","Metastasis_Stage","n()")


#######################
# 3. # Modeling Phase #
#######################

#Need Functions to plot
#Kaplan-Meier Survival plot Function
plot_surv = function(comp_group,plot_title,plot_data)
{
  print(survfit(formula = Surv_obj ~ comp_group, data = plot_data))
  ggsurvplot(
    surv_fit(formula = Surv_obj ~ comp_group, data = plot_data),
    conf.int = TRUE,
    conf.int.fill="strata",
    conf.int.alpha=0.2,
    palette = "Dark2",
    pval = TRUE,
    surv.median.line = "hv",
    legend="top",
    title=plot_title,
    break.time.by = 1
  )
}

#Kaplan-Meier Cummulative Hazard plot Function
plot_cumm_hazard = function(comp_group,plot_title,plot_data)
{
  ggsurvplot(
    surv_fit(formula = Surv_obj ~ comp_group, data = plot_data),
    conf.int = TRUE,
    conf.int.fill="strata",
    conf.int.alpha=0.2,
    palette = "Dark2",
    fun="cumhaz",
    pval = TRUE,
    legend = "top",
    title = plot_title,
    cumevents = TRUE,
    risk.table = TRUE
  )
}

#Parametric Fitting
fit_parametric = function(group)
{
  for(i in 1:length(distrib))
```

```
  {
    print(flexsurvreg(Surv_obj ~ group, data = temp, dist= distrib[i])$AIC)
  }
}

#########################################
# 3.1 # Comparison of different Cancers #
#########################################
Surv_obj = Surv(time = cancer_clin_data$Survival_Time, event = cancer_clin_data$Vi
tal_Status)

plot_surv(cancer_clin_data$Disease_Code,
          "Kaplan-Meier Survival plot of Different Cancer Types",
          cancer_clin_data)

plot_cumm_hazard(cancer_clin_data$Disease_Code,
                 "Kaplan-Meier Cummulative Hazard plot of Different Cancer Types",
                 cancer_clin_data)

survdiff(formula = Surv_obj ~ Disease_Code, data = cancer_clin_data)

vv = pairwise_survdiff(formula = Surv(time = Survival_Time,event = Vital_Status) ~
Disease_Code,
                       data = cancer_clin_data)

symnum(vv$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
       symbols = c("****", "***", "**", "*", "+", " "),
       abbr.colnames = FALSE, na = "")

####################
####################
## Breast Cancer ##
####################
####################

#####################################################################
# 3.2 # Survival of Breast Cancer Patients without any co-variates #
#####################################################################
Surv_obj = Surv(time = BRCA_data$Survival_Time, event = BRCA_data$Vital_Status)

fit_null = survfit(formula = Surv_obj ~ 1, data = BRCA_data)
print(fit_null)
ggsurvplot(
  surv_fit(formula = Surv_obj ~ 1, data = BRCA_data),
  conf.int = TRUE,
  conf.int.fill="strata",
  conf.int.alpha=0.2,
  palette = "Dark2",
  pval = TRUE,
  surv.median.line = "hv",
  legend="top",
```

```
  title="Kaplan-Meier Survival plot of Breast Cancer Patients without any covariat
es",
  break.time.by = 1,
  dist = "exponential"
)

#Life Table
summary(fit_null,times = c(.5*(1:40)))

ggsurvplot(
  surv_fit(formula = Surv_obj ~ 1, data = cancer_clin_data),
  pval = TRUE,
  conf.int = TRUE,
  conf.int.fill="strata",
  conf.int.alpha=0.2,
  fun="cumhaz",
  legend="top",
  palette = "Dark2",
  title="Kaplan-Meier Cummulative Hazard plot of Breast Cancer Patients without an
y covariates",
  risk.table=TRUE,
  cumevents=TRUE
)

################
# Age Category #
################
temp = BRCA_data %>% filter(!is.na(Age_Category))
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

plot_surv(temp$Age_Category,
          "Kaplan-Meier Survival plot of Different Age Category",
          temp)

plot_cumm_hazard(temp$Age_Category,
                 "Kaplan-Meier Cummulative Hazard plot of Different Age Category",
                 temp)

survdiff(formula = Surv_obj ~ Age_Category, data = temp)

vv=pairwise_survdiff(formula = Surv(Survival_Time,Vital_Status)~ Age_Category,data
= temp)

symnum(vv$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
       symbols = c("****", "***", "**", "*", "+", " "),
       abbr.colnames = FALSE, na = "")

#Cox-Proportional Hazard Rate
fit_coxph = coxph(formula = Surv_obj ~ Age_Category,data = temp)
ggcoxzph(cox.zph(fit_coxph),strata = TRUE)
summary(fit_coxph)
```

```
#Cox-PH for Continuous Age variate
fit_coxph = coxph(formula = Surv_obj ~ Age,data = temp)
ggcoxzph(cox.zph(fit_coxph),strata = TRUE)
summary(fit_coxph)

#Parametric Modelling
temp$Age_Category = relevel(temp$Age_Category, ref = "Middle Age")
distrib = c("exp","weibull","gamma","lnorm","llogis")
fit_parametric(temp$Age_Category)
summary(survreg(Surv_obj ~ Age_Category, data = temp, dist= "loglogistic"))

#################
# Race Category #
#################

temp=BRCA_data %>% filter(!is.na(Race))
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

plot_surv(temp$Race,
          "Kaplan-Meier Survival plot of Different Race Category",
          temp)

plot_cumm_hazard(temp$Race,
                 "Kaplan-Meier Cummulative Hazard plot of Different Race Category"
,
                 temp)

survdiff(formula = Surv_obj ~ Race, data = temp)

vv=pairwise_survdiff(formula = Surv(Survival_Time,Vital_Status)~ Race,data = temp)

symnum(vv$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
       symbols = c("****", "***", "**", "*", "+", " "),
       abbr.colnames = FALSE, na = "")

#Cox-Proportional Hazard Rate
fit_coxph = coxph(Surv_obj ~ Race,data = temp)
ggcoxzph(cox.zph(fit_coxph))
summary(fit_coxph)

#Parametric Modelling
distrib = c("exp","gamma","lnorm","llogis")
fit_parametric(temp$Race)
summary(survreg(Surv_obj ~ Race, data = temp, dist= "loglogistic"))

######################
# Ethnicity Category #
######################

temp = BRCA_data %>% filter(!is.na(Ethnicity))
```

```
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

plot_surv(temp$Ethnicity,
          "Kaplan-Meier Survival plot of Different Ethnicity Category",
          temp)

plot_cumm_hazard(temp$Ethnicity,
                 "Kaplan-Meier Cummulative Hazard plot of Different Ethnicity Cate
gory",
                 temp)

survdiff(formula = Surv_obj ~ Ethnicity, data = temp)

vv=pairwise_survdiff(formula = Surv(Survival_Time,Vital_Status)~ Ethnicity,data =
temp)

symnum(vv$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
       symbols = c("****", "***", "**", "*", "+", " "),
       abbr.colnames = FALSE, na = "")

#Cox-Proportional Hazard Rate
fit_coxph = coxph(formula = Surv_obj ~ Ethnicity,data = temp)
ggcoxzph(cox.zph(fit_coxph))

#Parametric Modelling
distrib = c("exp","weibull","gamma","lnorm","llogis")
fit_parametric(temp$Ethnicity)
summary(survreg(Surv_obj ~ Ethnicity, data = temp, dist= "loglogistic"))

#########################
# Therapy Type Category #
#########################
temp = BRCA_data %>% filter(!is.na(Therapy_Type))
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

plot_surv(temp$Therapy_Type,
          "Kaplan-Meier Survival plot of Different Therapy Types",
          temp)

plot_cumm_hazard(temp$Therapy_Type,
                 "Kaplan-Meier Cummulative Hazard plot of Different Therapy Types"
,
                 temp)

survdiff(formula = Surv_obj ~ Therapy_Type, data = temp)

vv=pairwise_survdiff(formula = Surv(Survival_Time,Vital_Status)~ Therapy_Type,data
= temp)

symnum(vv$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
       symbols = c("****", "***", "**", "*", "+", " "),
```

```
       abbr.colnames = FALSE, na = "")

#Cox-Proportional Hazard Rate
fit_coxph = coxph(Surv_obj ~ Therapy_Type,data = temp)
ggcoxzph(cox.zph(fit_coxph))
summary(fit_coxph)

#Parametric Modelling

distrib = c("exp","gamma","lnorm","llogis")
temp$Therapy_Type = relevel(temp$Therapy_Type, ref = "No Information")
fit_parametric(temp$Therapy_Type)
summary(survreg(Surv_obj ~ Therapy_Type, data = temp, dist= "loglogistic"))

#########################
# Cancer Stage Category #
#########################
temp = BRCA_data %>% filter(!is.na(Cancer_Stage))
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

plot_surv(temp$Cancer_Stage,
          "Kaplan-Meier Survival plot of Different Cancer Stages",
          temp)

plot_cumm_hazard(temp$Cancer_Stage,
                 "Kaplan-Meier Cummulative Hazard plot of Different Cancer Stages"
,
                 temp)

survdiff(formula = Surv_obj ~ Cancer_Stage, data = temp)

vv=pairwise_survdiff(formula = Surv(Survival_Time,Vital_Status)~ Cancer_Stage,data
= temp)

symnum(vv$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
       symbols = c("****", "***", "**", "*", "+", " "),
       abbr.colnames = FALSE, na = "")

#Cox-Proportional Hazard Rate
fit_coxph = coxph(Surv_obj ~ Cancer_Stage,data = temp)
ggcoxzph(cox.zph(fit_coxph))
summary(fit_coxph)

#Parametric Modelling
distrib = c("exp","weibull","gamma","lnorm","llogis")
fit_parametric(temp$Cancer_Stage)
summary(survreg(Surv_obj ~ Cancer_Stage, data = temp, dist= "loglogistic"))

########################
# Tumor Stage Category #
########################
```

```
temp = BRCA_data %>% filter(!is.na(Tumor_Stage))
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

plot_surv(temp$Tumor_Stage,
          "Kaplan-Meier Survival plot of Different Tumor Stages",
          temp)

plot_cumm_hazard(temp$Tumor_Stage,
                 "Kaplan-Meier Cummulative Hazard plot of Different Tumor Stages",
                 temp)

survdiff(formula = Surv_obj ~ Tumor_Stage, data = temp)

#Cox-Proportional Hazard Rate
fit_coxph = coxph(Surv_obj ~ Tumor_Stage,data = temp)
ggcoxzph(cox.zph(fit_coxph))

#Parametric Modelling
temp$Tumor_Stage = as.factor(temp$Tumor_Stage)

temp$Tumor_Stage = relevel(temp$Tumor_Stage,ref = "Tumor Stage 4")
distrib = c("exp","weibull","gamma","lnorm","llogis")
fit_parametric(temp$Tumor_Stage)
summary(survreg(Surv_obj ~ Tumor_Stage, data = temp, dist= "loglogistic"))

#############################
# Lymph Node Stage Category #
#############################
temp = BRCA_data %>% filter(!is.na(Lymph__node_Stage))
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

plot_surv(temp$Lymph__node_Stage,
          "Kaplan-Meier Survival plot of Different Lymph Node Stages",
          temp)

plot_cumm_hazard(temp$Lymph__node_Stage,
                 "Kaplan-Meier Cummulative Hazard plot of Different Lymph Node Sta
ges",
                 temp)

survdiff(formula = Surv_obj ~ Lymph__node_Stage, data = temp)

vv=pairwise_survdiff(formula = Surv(Survival_Time,Vital_Status)~ Lymph__node_Stage
,data = temp)

symnum(vv$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
       symbols = c("****", "***", "**", "*", "+", " "),
       abbr.colnames = FALSE, na = "")

#Cox-Proportional Hazard Rate
fit_coxph = coxph(Surv_obj ~ Lymph__node_Stage,data = temp)
```

```
ggcoxzph(cox.zph(fit_coxph))
summary(fit_coxph)

#Parametric Modelling
distrib = c("exp","weibull","gamma","lnorm","llogis")
fit_parametric(temp$Lymph__node_Stage)
summary(survreg(Surv_obj ~ Lymph__node_Stage, data = temp, dist= "loglogistic"))

#############################
# Metastasis Stage Category #
#############################
temp = BRCA_data %>% filter(!is.na(Metastasis_Stage))
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

plot_surv(temp$Metastasis_Stage,
          "Kaplan-Meier Survival plot of Different Metastasis Stages",
          temp)

plot_cumm_hazard(temp$Metastasis_Stage,
                 "Kaplan-Meier Cummulative Hazard plot of Different Metastasis Sta
ges",temp)

survdiff(formula = Surv_obj ~ Metastasis_Stage, data = temp)

vv=pairwise_survdiff(formula = Surv(Survival_Time,Vital_Status)~ Metastasis_Stage,
data = temp)

symnum(vv$p.value, cutpoints = c(0, 0.0001, 0.001, 0.01, 0.05, 0.1, 1),
       symbols = c("****", "***", "**", "*", "+", " "),
       abbr.colnames = FALSE, na = "")

#Cox-Proportional Hazard Rate
fit_coxph = coxph(Surv_obj ~ Metastasis_Stage,data = temp)
ggcoxzph(cox.zph(fit_coxph))

#Parametric Modelling
distrib = c("exp","weibull","gamma","lnorm","llogis")
fit_parametric(temp$Metastasis_Stage)
summary(survreg(Surv_obj ~ Metastasis_Stage, data = temp, dist= "loglogistic"))

##################
## Cox-PH Model ##
##################
temp = BRCA_data %>% filter(!is.na(Age)) %>% filter(!is.na(Therapy_Type)) %>% filt
er(!is.na(Lymph__node_Stage))
Surv_obj = Surv(time = temp$Survival_Time, event = temp$Vital_Status)

fit_coxph = coxph(Surv_obj ~ Age + Therapy_Type + Lymph__node_Stage,data = temp)
print(cox.zph(fit_coxph))
ggcoxzph(cox.zph(fit_coxph))
summary(fit_coxph)
```

9. References

Zhang Z. Parametric regression model for survival data: Weibull regression model as an example. Ann Transl Med. 2016;4(24):484. doi:10.21037/atm.2016.08.45

Jackson CH. flexsurv: A Platform for Parametric Survival Modeling in R. J Stat Softw. 2016;70:i08. doi:10.18637/jss.v070.i08

Zhang, Zhongheng. (2016). Semi-parametric regression model for survival data: Graphical visualization with R. Annals of Translational Medicine. 4. 461-461. 10.21037/atm.2016.08.61.

Zhang Z. Statistical description for survival data. Ann Transl Med 2016;4(20):401. doi: 10.21037/atm.2016.07.17

Cancer.org - https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html

Breastcancer.org - https://www.breastcancer.org/symptoms/diagnosis/staging

GDC Webisite - https://portal.gdc.cancer.gov/