1 Introduction

2 Research Questions

3 Survival Analysis Basics

4 R Functions for Survival Analysis

5 Data Source and Description

6 Exploratory Data Analysis

7 Data Analysis

8 Summary of Results

9 Answers to Research Questions:

10 References

Survival Analysis of Breast Cancer Data from the TCGA Dataset

Ramaa Nathan

6/25/2019

1 Introduction

Survival Analysis is a branch of statistics to study the expected duration of time until one or more events occur, such as death in biological systems, failure in mechanical systems, loan performance in economic systems, time to retirement, time to finding a job in etc. In case of cancer studies, one of the primary objectives is to assess the time to an event of interest like relapse of cancer, death, etc.

Survival Analysis is especially helpful in analyzing these studies when one or more of the cohorts do not experience the event and are considered censored for various reasons like death due to a different cause, loss-to-follow-up, end of study, etc. The basic quantity used to describe time-to-event data is the survival function which is the probability of surviving beyond time x.

The survival function can be modeled using parametric methods (Exponential, Weibull, etc), semi-parametric methods (Cox proportional hazards model), and non-parametric methods (Kaplan Meier model). The difference in survival times due to different treatment groups can also be compared using Logrank tests. The Cox proportional hazards model is useful in modeling the survival function in the presence of covariates.

The Cancer Genome Atlas (TCGA) Program (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga), a joint effort between the National Cancer Institute and the Human Genome Research Institute, provides publicly-available clinical and high-throughput genomic data for thirty-three different types of cancers. This rich data source is widely used by researchers and has led to vast improvements in diagnosing, treating, and preventing cancer. The TCGA dataset will be used in this study to do survival analysis of breast cancer data.

2 Research Questions

Here are a few questions that we could answer with this study.

1. What is the probability of survival for breast cancer?
2. How do different cancers (breast, ovarian, lung) affect survival rates?
3. What are the important factors that influence estimation of survival rate for Breast Cancer?
4. What are the effects of each factor on survival?
5. What is the probability of survival for breast cancer when other clinical covariates are considered?

3 Survival Analysis Basics

Before we start analysing the data, lets first try to understand the basic terminologies related to survival analysis.

3.1 Terminologies

3.1.1 Event Times

Event times often are useful endpoints in clinical trials. Examples include survival time from onset of diagnosis, time until progression from one stage of disease to another, and time from surgery until hospital discharge. In each case, time is measured from study entry until the event occurs.

3.1.2 Censoring

With an endpoint that is based on an event time, there always is the chance of censoring. An event time is censored if there is some amount of follow-up on a subject and the event is not observed during the study period. There are different types of censoring **???**

3.1.2.1 Right Censoring

Right censoring occurs when an event is not observed because of loss-to-follow-up, death from a cause other than the trial endpoint, study termination, and other reasons unrelated to the endpoint of interest. This occurs frequently in studies of survival. There are three types of right censoring:

1. Type I censoring occurs when all subjects are scheduled to begin the study at the same time and end the study at the same time. This type of censoring is common in laboratory animal experiments, but unlikely in human trials.
2. Type II censoring occurs when all subjects begin the study at the same time and the study is terminated when a predetermined proportion of subjects have experienced the event.
3. Type III censoring occurs when the censoring is random, which is the case in clinical trials because of staggered entry (not every patient enters the study on the first day) and unequal follow-up on subjects.

3.1.2.2 Left Censoring

Left censoring occurs when the initiation time for the subject, such as time of diagnosis, is unknown.

3.1.2.3 Interval Censoring

Interval censoring occurs when the subject is not followed for a period of time during the trial and it is unknown if the event occurred during that period.

3.1.3 Survival Function

The survival function is the probability of surviving beyond time t or the probability of experiencing the event beyond time t. The survival function takes value 1 at the origin and 0 at infinity.

$$S(t) = P(T > t)$$

If f(t) is the probability density function (pdf) that describes the time-to-event, and F(t) is the corresponding cumulative distributive function, then

$$F(t) = \int_0^t f(x)\,dx$$

$$S(t) = 1 - F(t) = \int_t^\infty f(x)\,dx$$

3.1.4 Hazard Rate or Hazard Function

The hazard function, h(x) is defined as the instantaneous risk of the event or the probability that if a person survives to t, they will experience the event in the next instant. The hazard function or hazard rate can be considered to be the slope of the survival function.

$$h(t) = \lim_{\Delta t -> 0} \frac{P[t \leq T < t + \Delta\, t | T \geq t]}{\Delta t}$$

The hazard function can also be expressed as the ratio of the probability density function to the survival function

$$h(t) = \frac{f(t)}{S(t)}$$

3.1.5 Cumulative hazard

Cumulative hazard is the accumulation of hazard rate over time. It is not a probability and is a measure of the risk. The greater the value of H(t), the greater the risk of failure by time t. When the distribution is continuous, the cumulative hazard is the integral of the hazard rate.

$$H(t) = \int_0^t h(x)dx = -log(S(t))$$

3.1.6 Hazard Ratio

Hazard ratio is defined as the ratio of of two hazard functions, corresponding to two treatment groups

$$\Lambda = \frac{h_1(t)}{h_2(t)}$$

3.1.7 Proportional Hazard

The hazard ratio can be used to compare two treatment groups. When the hazard ratio is constant and independent of time, the hazards for the two treatment groups are said to be proportional. Basically, the relative risk of the event is constant over time.

3.2 Modeling Survival Function

The survival function can be modeled using non-parametric, parametric, and semi-parametric techniques.

3.2.1 Non-Parametric Methods

In a non-parametric method, we do not assume any distribution for the survival data and use only empirical information from the data. Kaplan Meier and the Cutler-Ederer are two non-parametric methods used to model the survival curve. Here, we will consider the most commonly used Kaplan Meier method.

3.2.1.1 Kaplan Meier

The Kaplan Meier survival curve is a non-parametric technique for estimating the probability of survival, even in the presence of censoring. In this model, there is the notion of a risk set, which is the set of all individuals who are at risk to have an event at time t. This includes individuals who are known to be alive at time t and those who have the actual event at time t.

In the modeling process, the actual failure times (event times) are first ordered in an increasing order. At each event time t_k, the number of subjects still at risk (n_k), the number of events (d_k), number of censored (lost to follow-up) subjects since the last event time ($n_k - d_k$) are recorded. The risk set does not include the subjects lost to follow-up. The Kaplan Meier Survival probability at time t_k utilizes conditional probability - the probability of surviving at time t, given that the person has survived upto time t.

$$S(t) = \begin{cases} 1 & t_0 \leq t \leq t_1 \\ \prod_{i=1}^{i=k}(\frac{n_i - d_i}{n_i}) & t_k \leq t \leq t_{k+1}, k = 1, 2, \ldots, K \end{cases}$$

Tha basic assumptions in a Kaplan Meier model are:

```
1. The censoring is independent of prognosis
2. The survival probabilities are the same for all subjects recruited at any time in the study
3. Events happened at the time specified.
```

3.2.1.2 Comparison of Survival Curves using Log-Rank

An important component of survival analysis is to compare the survival curves for two different groups. Even though the the survival rate at a given time t_x may be the same for two groups, the survival rates may vary at all other times. For example, the survival rate may steadily decrease for one group but may initially decrease rapidly for the other group and then be steady till time t_x.

Here, we will compare the non-parametric Kaplan Meier survival curves of two different groups using the Mantel-Haenszel test that is also called the log rank test. In this test, the following hypothesis is being tested.

Null Hypothesis: There is no difference in the survivial functions for the different groups.
Alternate Hypothesis: The survival functions of at aleast one of the groups is different.

For example, consider two groups - intervention and control. In the procedure described by Conchran, Mantel and Haenszel, the following 2x2 table is created at each time instant t_j that an event occurs.

Types	Death at time tj	Survivors at time tj	At risk prior to time tj
Intervention	a_j	b_j	$a_j + b_j$
Control	c_j	d_j	$c_j + d_j$
Total	$a_j + c_j$	$b_j + d_j$	n_j

a_j represents the observed number of deaths in the intervention group at time t_j and c_j represents the observed number of deaths in the control group at time t_j. Of the total n_j number of subjects at risk at time t_j, $a_j + b_j$ represents the total number of subjects at risk in the intervention group and $c_j + d_j$ represents the total number of subjects at risk in the control group.

The expected number of deaths in the intervention group can be expressed as

$$E(a_j) = \frac{(a_j + c_j)(b_j + d_j)}{n_j}$$

The variance of the observed number of deaths in the intervention group is given by

$$V(a_j) = \frac{(a_j + c_j)(b_j + d_j)(a_j + b_j)(c_j + d_j))}{n_j^2(n_j - 1)}$$

The MH statistic or the log-rank statistic to test the hypothesis is given by

$$MH = \frac{(O - E)^2}{V} = \frac{(\Sigma_{j=1}^{K} a_j - E(a_j))^2}{\Sigma_{j=1}^{K} V(a_j)}$$

The MH statistic has approximately a chi-square distribution with one degree of freedom (when only two curves are being compared). An asymptotic approximation is the Z_{MH} , the signed square root of MH and with a standard normal distribution.

$$Z_{MH} = \frac{\Sigma_{j=1}^{K} a_j - E(a_j)}{\sqrt{(\Sigma_{j=1}^{K} V(a_j))}}$$

The log rank test is a non-parametric test. So there are no assumptions made on the distribution of the survival curves. The same assumptions made for the Kaplan Meir test hold good for this test.

3.2.2 Parametric Methods

In parametric methods, the common distributions used to model the survival curve are Exponential Distribution, Weibull Distribution, and the Log-logistic distribution. We will consider the Exponential and Weibull Distributions here.

3.2.2.1 Exponential Distribution

$$
\begin{aligned}
pdf : f(t) &= \lambda e^{-\lambda t} \\
cdf : F(t) &= 1 - e^{-\lambda t} \\
Survival\ Function : S(t) &= 1 - F(t) = e^{-\lambda t} \\
Hazard\ Function : h(t) &= \frac{f(t)}{S(t)} = \lambda \\
Hazard\ Ratio &= \frac{\lambda_1}{\lambda_2} \\
Cumulative\ Hazard : H(t) &= -logS(t) = \lambda t
\end{aligned}
$$

In the case of an exponential distribution, the hazard function or the hazard rate is a constant λ and the hazard ratio is proportional as it is independent of time.

3.2.2.2 Weibull Distribution

$$
\begin{aligned}
pdf : f(t) &= p\lambda t^{p-1} e^{-(\lambda t^p)} \\
cdf : F(t) &= 1 - e^{-(\lambda t^p)} \\
Survival\ Function : S(t) &= 1 - F(t) = e^{-(\lambda t^p)} \\
Hazard\ Function : h(t) &= \frac{f(t)}{S(t)} = p\lambda t^{p-1} \\
Cumulative\ Hazard : H(t) &= (\lambda t^p)
\end{aligned}
$$

Here, $\lambda > 0$ is the scale parameter and $p > 0$ is the shape parameter. When the shape parameter equals 1, the Weibull distribution reduces to the exponential distribution. With a weibull distribution, the hazard function is not a constant and is dependent on time.

3.2.2.3 Understanding Parametric Modeling in R

In R, the survival::survreg() function is used for parametric modeling of the survival data. The survreg() function uses the accelerated failure-time (AFT) model, which assumes that the log of survival time, log T, can be expressed as a linear combination of μ the mean survival time, the covariates and parameters $\gamma_i z_i$ and an error term σW, where W is any distribution.

$$logT = \mu + \Sigma_i \alpha_i z_i + \sigma W$$

Recall that in an exponential distribution, the pdf $f(t) = \lambda e^{-\lambda t}$, λ is the mean number of events within an unit time and $1/\lambda$ is the mean waiting time for the first event to occur. Mapping this to the AFT model above, μ. the mean survival time is nothing but the mean waiting time. In the output of the survreg function, the value of the intercept corresponds to μ in the model. So, MLE of the hazard rate, λ can be computed as the $exp(-1/\mu)$ or $exp(-1/Intercept)$. The coefficients are logarithms of ratios of survival times, so a positive coefficient means longer survival.

3.2.3 Semi-Parametric Cox Proportional Hazard Regression Model

The Cox Regression Model is used to model the hazard at time t in the presence of multiple covariates, each of which could be categorical or quantitative. The Cox model is similar to the exponential model where the survival time is given by $S(t) = e^{\lambda t)}$. But the hazard rate λ is now considered to be a linear combination of several covariates $Z = Z_1, Z_2, .., Z_p$ and so

$$\lambda(Z_1, Z_2, .., Z_p) = \beta_1 Z_1 + \beta_2 Z_2 + .. + \beta_p Z_p$$

The Cox regression model can then be expressed as

$$h(t|Z) = h_0(t) exp(\Sigma_{k=1}^{p} \beta_k Z_k)$$

where $h(t|Z)$ is the hazard at time t for an individual with covariates Z and $h_0(t) = h(t|Z = 0)$ is the baseline hazard rate.

The Cox model is semi-parametric, containing both parametric and non-parametric components.

1. The $h_0(t)$ is the non-parametric component and can take any form as along as $h_0(t) \geq 0$
2. $exp(\Sigma_{k=1}^{p} \beta_k Z_k)$ is the parametric component.

For example, suppose we have three predictors - therapy, age and race. Then

$$h(t) = h_0(t) exp(\beta_1 therapy + \beta_2 age + \beta_3 race)$$

Now given two individuals with same covariates, ie. of the same age and race, but that the first individual gets treatment B (corresponding indicator variable has value 1) and second individual gets treatment A (corresponding indicator variable has value 0). Then the hazards for the two individuals are:

$$
\begin{aligned}
h_1(t) &= h_0(t) exp(\beta_1 + \beta_2 age + \beta_3 race) \\
h_2(t) &= h_0(t) exp(\beta_2 age + \beta_3 race) \\
HazardRatio = \frac{h_1(t)}{h_2(t)} &= \frac{h_0(t) exp(\beta_1 + \beta_2 age + \beta_3 race)}{h_0(t) exp(\beta_2 age + \beta_3 race)} = exp(\beta_1) \\
log(HazardRatio) &= \beta_1
\end{aligned}
$$

So, the coefficient β_1 is the log of the hazard ratio and $exp(\beta_1)$ is the hazard ratio for the individual on treatment B compared to treatment A, when the race and age

covariates are the same for both individuals. The values of $exp(\beta 1)$ provide the following interpretations,

```
* $exp(\beta_1)$ > 1 indicates higher hazard or lower survival rate compared to the base hazard function.
* $exp(\beta_1)$ < 1 indicates lower hazard or higher survival rate
* $exp(\beta_1)$ = 1 indicates no association
```

In R, we use the coxph function to model the Cox PH model. It provides the following output values:

```
 1. coef = the estimate of $\beta_i$
 2. exp(coef) - the estimate of $exp(\beta_i)$
 3. se(coef) - the standard error of the estimate of $\beta_i$
 4. z = $\frac{z}{se(coef)}$ = the Wald statistic for testing the null hypothesis that $\beta_i = 0$ assuming that z follows a stand
ard normal distribution
 5. p = two sided p-value
```

One of the basic assumptions of the Cox Proportional Hazards Model is that the hazards are proportional. This can be tested by checking that the Schoenfeld residuals exhibit a random pattern against time. In R, cox.zph function in the survival package and ggcoxzph from the survminer package can be used to check for the plot of these residuals.

4 R Functions for Survival Analysis

The survival analysis in this study has been done in R and here is a list of the important functions used in the analysis.

R Functions for Survival Analysis		
Purpose	**Package::Function**	**Package::Graphical Wrapper**
Extract Survival Data from TCGA	rtcga::survivalTCGA	
Create Survival Object	survival::Surv()	
Fit a Kaplan Meier Curve	survival::survfit	survminer::ggsurvplot()
Compare Kaplan Meier Curves using logrank	survival::survdiff()	survminer::ggsurvplot()
Fit a parametric model	survival::survreg	
Fit Cox Proportional Hazards Model	survival::coxph()	survminer::ggforest()
Test for Proportional Hazards	survival::cox.zph()	survminer::ggcoxzph()
Display Adjusted Survival Curves for Cox Proportional Hazards Model for a Factor		survminer::ggcoxadjustedcurves
Split the survival data set at specific cut times to accommodate the time-dependent covaraites for Cox	survival::survSplit	
Convert Weibull restuls to an easy interpretable form	SurvRegCensCov::ConvertWeibull	

5 Data Source and Description

The Cancer Genome Atlas (TCGA) Program [1] provides publicly-available clinical and high-throughput genomic data for thirty-three different types of cancers. For this study of survival analysis of Breast Cancer, we use the Breast Cancer (BRCA) clinical data that is readily available as BRCA.clinical. This dataset has 3703 columns from which we pick the following columns containing demographic and cancer stage information as important predictors of survival analysis.

ColumnName	DataType	Description
Gender	categorical	Gender
Race	categorical	Race
Ethnicity	categorical	Ethnicity
Age	integer	Age at first diagnosis
Vital Status	binary	Vital Status (1 - dead (event), 0 - alive/censored)
Days to Death	integer	Number of days to death from first diagnosis
Days to Followup	integer	Number of days to last follow-up from first diagnosis
Therapy type	categorical	Therapy Type (Chemo, Harmone, Immuno, etc.
Pathologic Stage	categorical	Cancer stage - based on T,M, and N labeling
Pathology T	categorical	Tumor (T) stage describing size and location of tumor

The descriptions of the pathology stages as given by NCIthesaurus

Pathology N	categorical	Lymph (N) nodes status describing if cancer has spread into nearby lymph nodes
Pathology M	categorical	Metastasis (M) status describing if cancer has spread to other parts of the body

(https://ncit.nci.nih.gov/ncitbrowser/pages/home.jsf?version=19.05d)

1. Pathologic Stage: This is the classification of cancer stages and is based on the T,M,and N staging
    a. Stage 1: Invasive cancer confined to the original anatomic site of growth without lymph node involvement
    b. Stage 2: Invasive cancer more extensive than stage I, usually involving local lymph nodes without spread to distant anatomic sites.
    c. Stage 3: Locally advanced cancer that has spread to nearby organs but not to distant anatomic sites.
    d. Stage 4: Cancer that has spread to distant anatomic sites beyond its original site of growth.
2. Tumor Stages:
    a. Stage T1: A clinical and/or pathologic primary tumor TNM finding indicating that the cancer is limited to the site of growth.
    b. Stage T2: For breast cancer it refers to primary tumor that is more than 2.0 cm, but not more than 5.0 cm in greatest dimension
    c. Stage T3: A clinical and/or pathologic primary tumor TNM finding usually indicating that the cancer is locally invasive, without infiltration of adjacent structures.
    d. Stage T4: A clinical and/or pathologic primary tumor TNM finding indicating direct invasion of adjacent structures by cancer.
    e. Stage TX: A primary tumor TNM finding indicating that the status of the primary tumor cannot be assessed.
3. Metastatis Stages:
    a. Stage M0: A distant metastasis TNM finding indicating that there is no evidence of distant metastasis.
    b. Stage CM0: Breast cancer without clinical or radiographic evidence of distant metastases.
    c. Stage M1: A clinical and/or pathologic distant metastasis TNM finding indicating the spread of cancer to distant anatomic sites
    d. Stage MX: A distant metastasis TNM finding indicating that the status of distant metastasis cannot be assessed.
4. Lymph Nodes Stages:
    a. Stage N0: A regional lymph node TNM finding indicating that there is no evidence of regional lymph node metastasis.
    b. Stage N1: For breast cancer it refers to micrometastases or metastases in 1-3 axillary lymph nodes;
    c. Stage N2: For breast cancer it refers to metastases in 4-9 axillary lymph nodes;
    d. Stage N3: for breast cancer it refers to metastases in 10 or more axillary lymph nodes;
    e. Stage NX: A regional lymph node TNM finding indicating that the status of regional lymph nodes cannot be assessed.

6 Exploratory Data Analysis

The clinical data set from the The Cancer Genome Atlas (TCGA) Program is a snapshot of the data from 2015-11-01 and is used here for studying survival analysis.

6.1 Data Extraction

The RTCGA package in R is used for extracting the clinical data for the Breast Invasive Carcinoma Clinical Data (BRCA). In addition, the survival and survminer packages in R are used for the analysis.

The survivalTCGA function in the RTCGA package is used to extract the relevant columns. This function also uses the vital status variable that indicates if the observation was an event or a censor and combines the number of days to death from first diagnosis and number of days to last follow-up from first diagnosis into a new “times” variable.

6.2 Checks for Missing and Invalid Data

The extracted data is first checked for any missing data or invalid data. There are two observations that have negative values for the time to event that are filtered out during the data cleanup and transformation step.

The data is generally clean with only some of the demographic information like race and ethnicity missing for a few of the observations. We also find that more than 40% of the rows have NAs in one or more columns. So, we will filter out the missing data, as needed during the analysis.

6.3 Data Cleaning and Transformation

Next, the following cleaning and transformations are applied to the TCCGA BRCA.clinical data to get a clean and compact dataset:

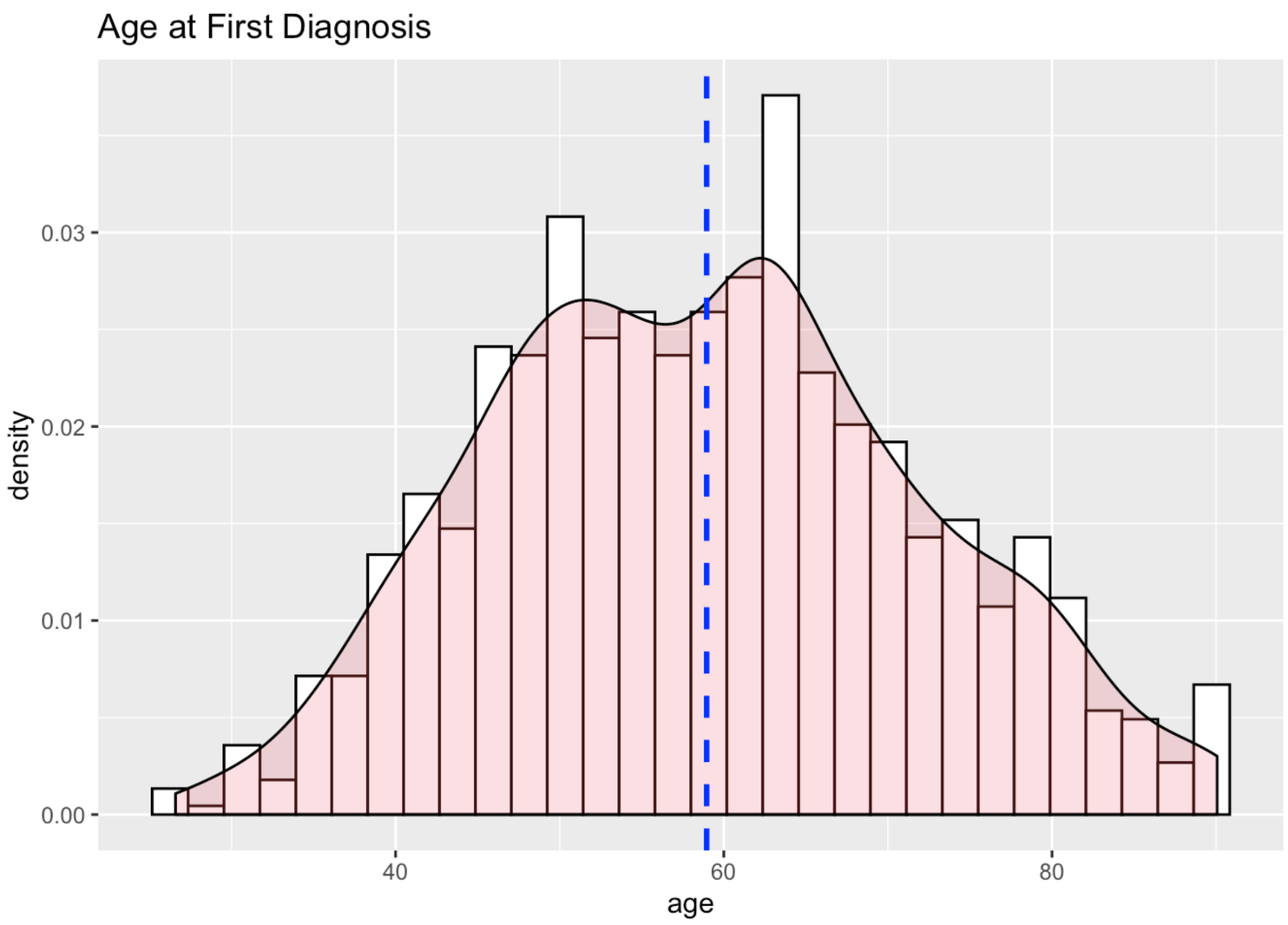
1. Rename the long variable names to short names.
2. Filter out the 12 observations corresponding to males diagnosed with breast cancer.
3. Filter out the 2 observations with negative “times” value.
4. Create a “age” variable that contains the number of days at first diagnosis to age in years at first diagnosis.
5. Create a “years_to_event” variable which is the “times”" variable converted from days to years.
6. Data in the pathology columns contain information on both stage and sub-stage. Transform the data to only contain the high level stage information.
7. Modify the “therapy_type”" to contain three types - chemotherapy, hormone therapy and Other (lump all the other infrequent types into Other)
8. Modify the “race”" to contain three types - black or african american and white (lump the other two types into Other)

6.4 Data Visualizations

6.4.1 Histograms

Age at First Diagnosis

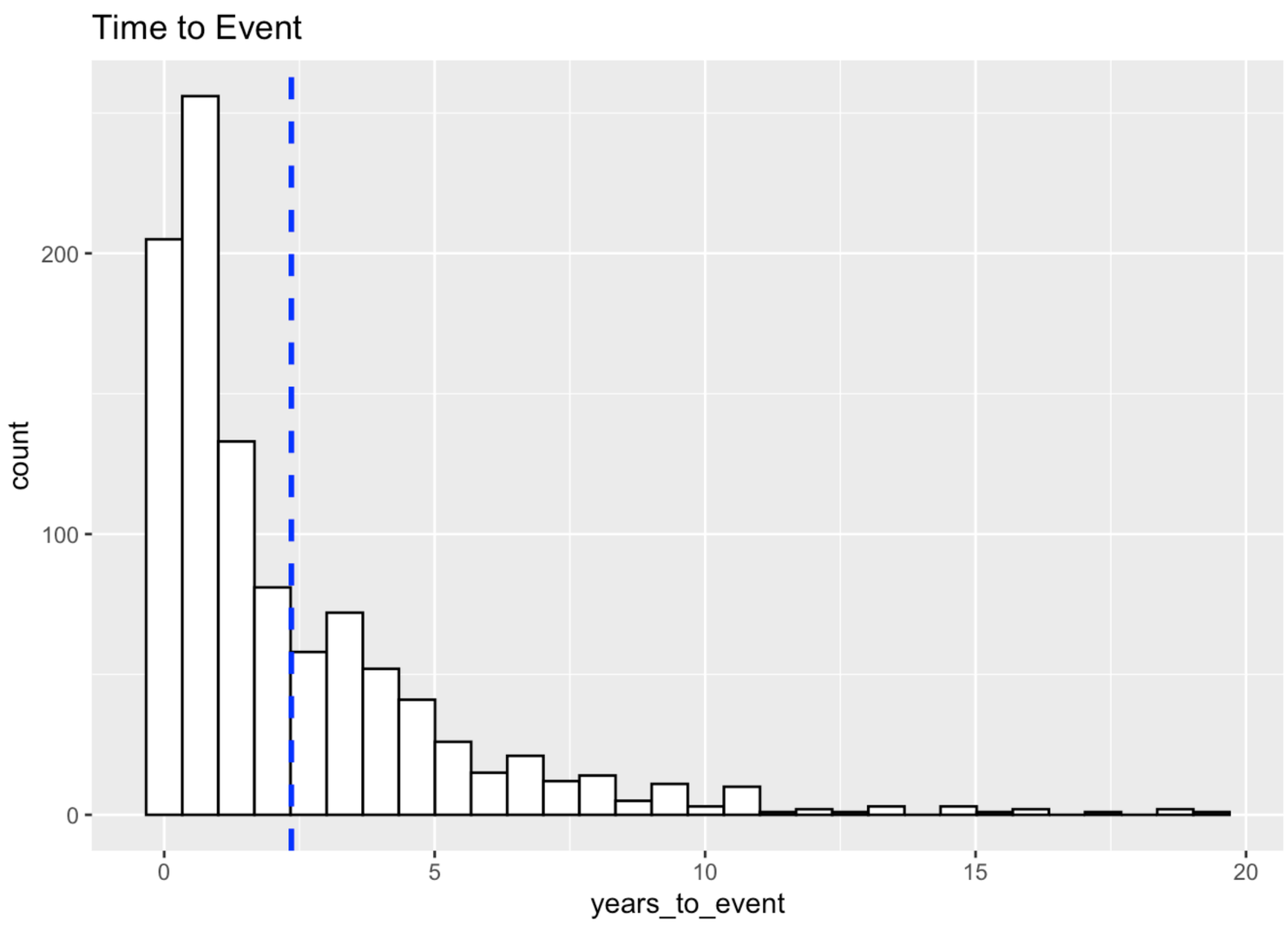
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The average age at first diagnosis seems to 59 years and the distribution of age is mostly symmetrical with a small bimodal effect.

Time to Event

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

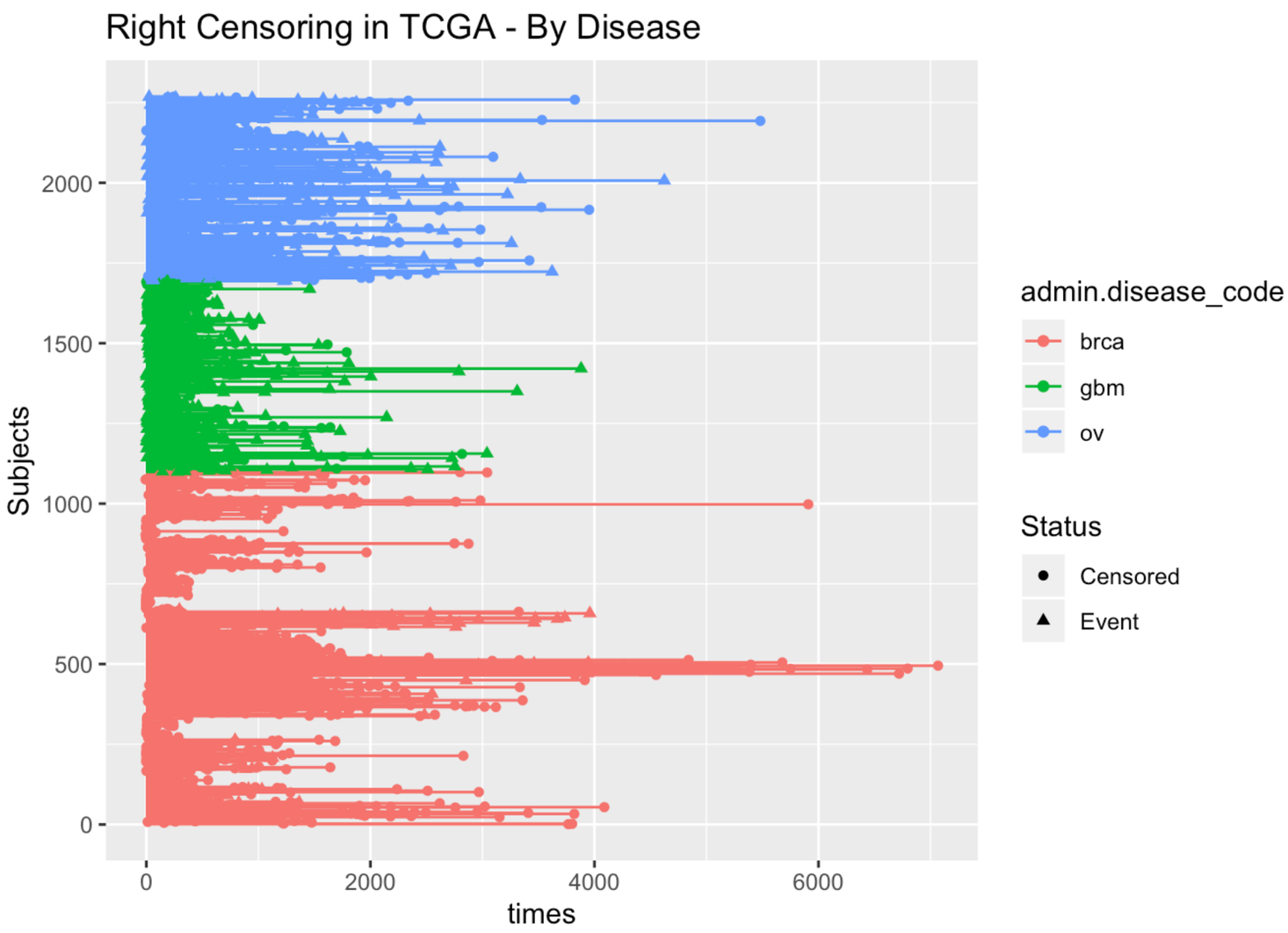



```
##       age
##  Min.   :26.59
##  1st Qu.:49.28
##  Median :58.94
##  Mean   :58.95
##  3rd Qu.:68.00
##  Max.   :90.06
##  NA's   :9
```

Without considering censoring the mean survival time seems to be less than 2.5 years. This is not helpful information as censoring information is an important component of survival data. The TCGA clinical data set is a survival dataset containing right censor information. We will first visualize the distribution of the right censor data by different categories. The Censoring Event plots here were inspired by a workshop on Survival Analysis (https://github.com/emsweene/New_York_R_Survival_Analysis_Workshop).

6.4.2 Censoring and Event Plots by Disease

```
##                   patient.vital_status
## admin.disease_code    0    1  Sum
##               brca  994  104 1098
##               gbm   149  446  595
##               ov    279  297  576
##               Sum  1422  847 2269
```

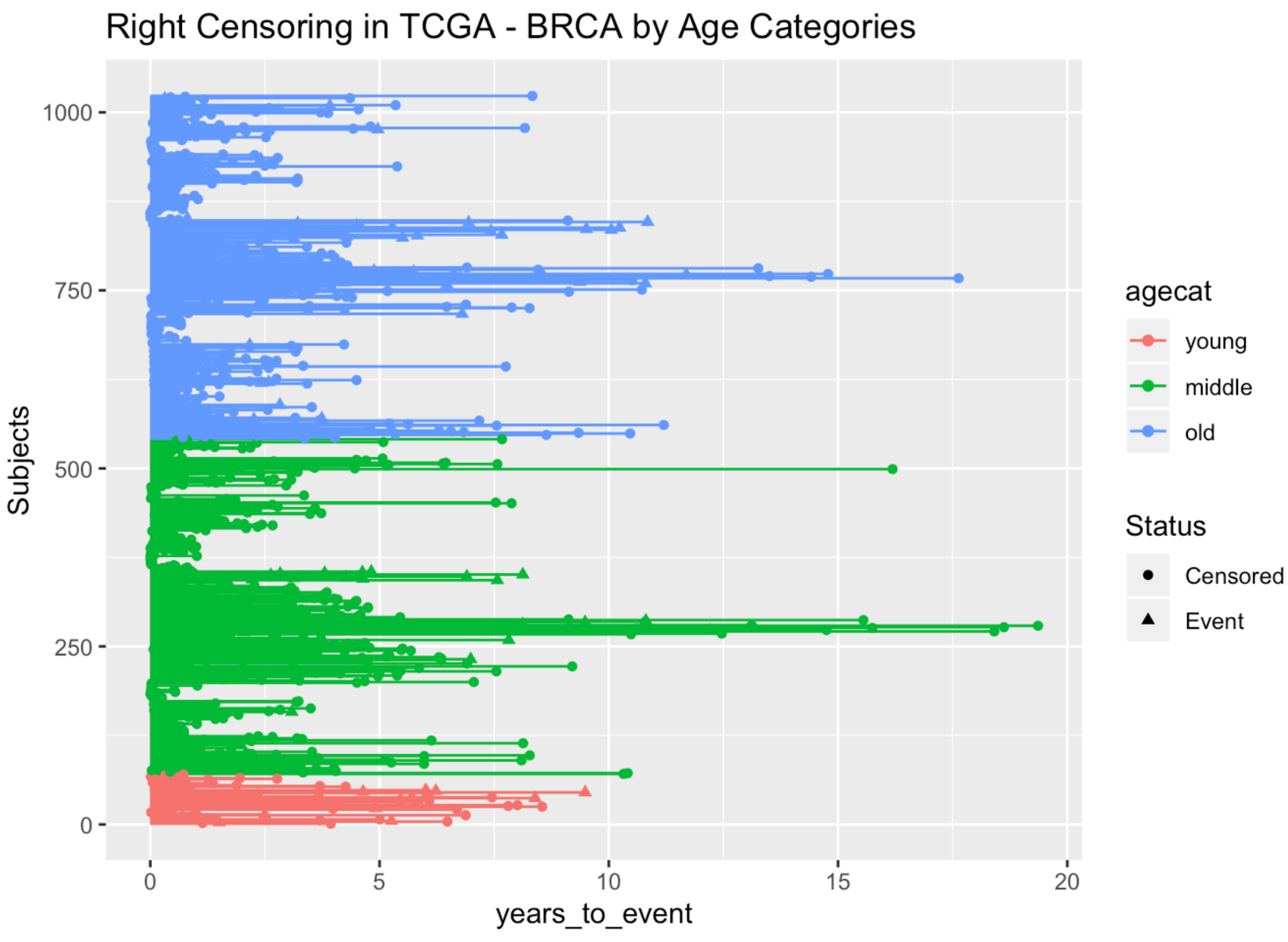


From the contingency table and the plots comparing the diseases of BRCA (breast cancer), GBM (Glioblastoma Multiforme) and OV (Ovarian Cancer), it can be observed that less that almost 50% of the cases are for Breast Cancer and the rest are almost equally split between ovarian and GBM. In these, 10% of the subjects with Breast cancer, 75% of the subjects with GBM and 50% of the subjects with ovarian cancer did not survive (had events).

6.4.3 Censoring and Event Plots by Age Category

,

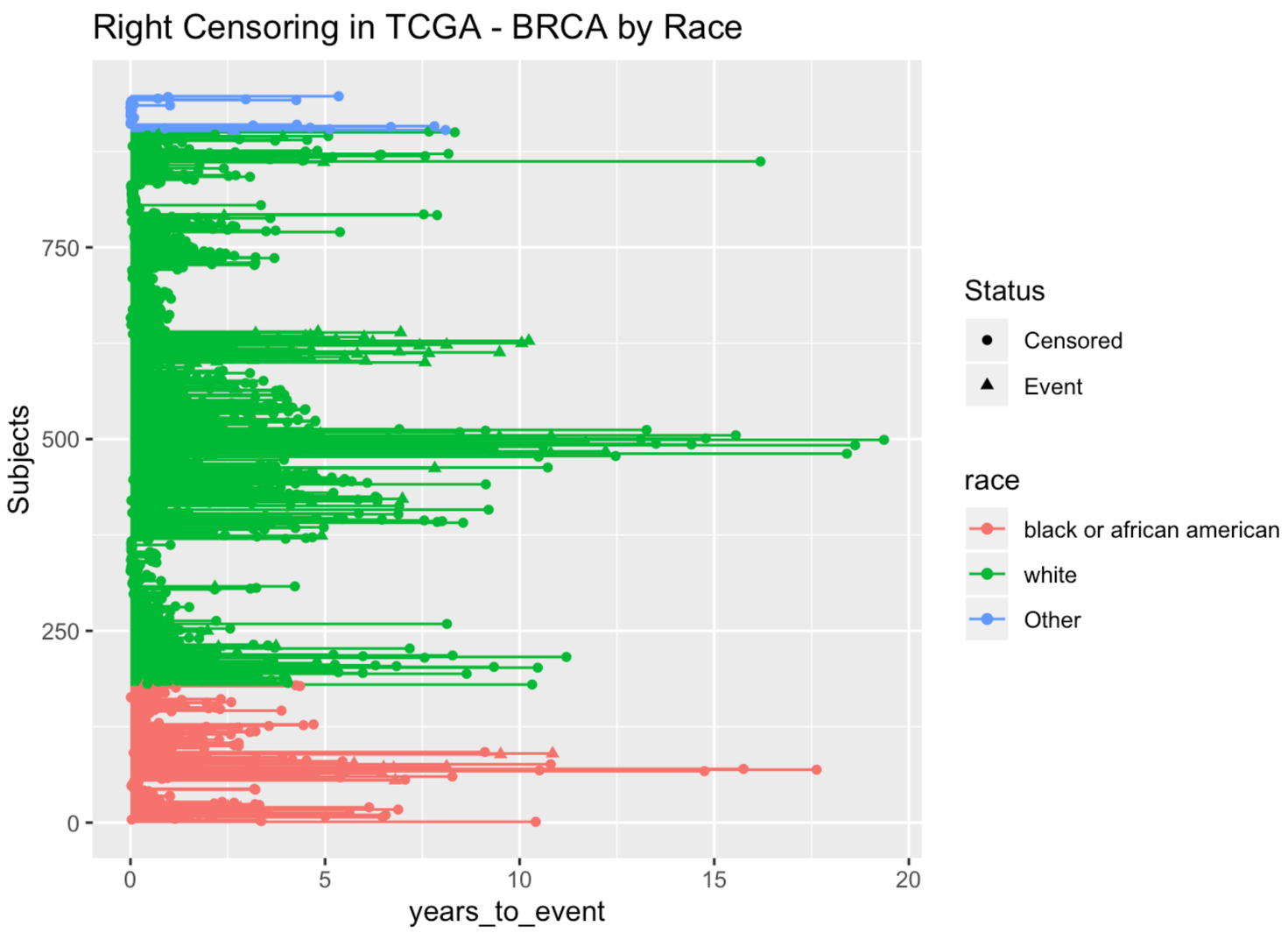
```
##          vital_status
## agecat       0    1  Sum
##   young     59   11   70
##   middle   436   35  471
##   old      424   58  482
##   Sum      919  104 1023
```



From the contingency table and the plots, it can be observed that less thatn 10% of the subjects are less than 40 years old. The number of subjects in the other two groups are almost equally distributed of which more number of events were observed in the older age group.

6.4.4 Censoring and Event Plots by Race

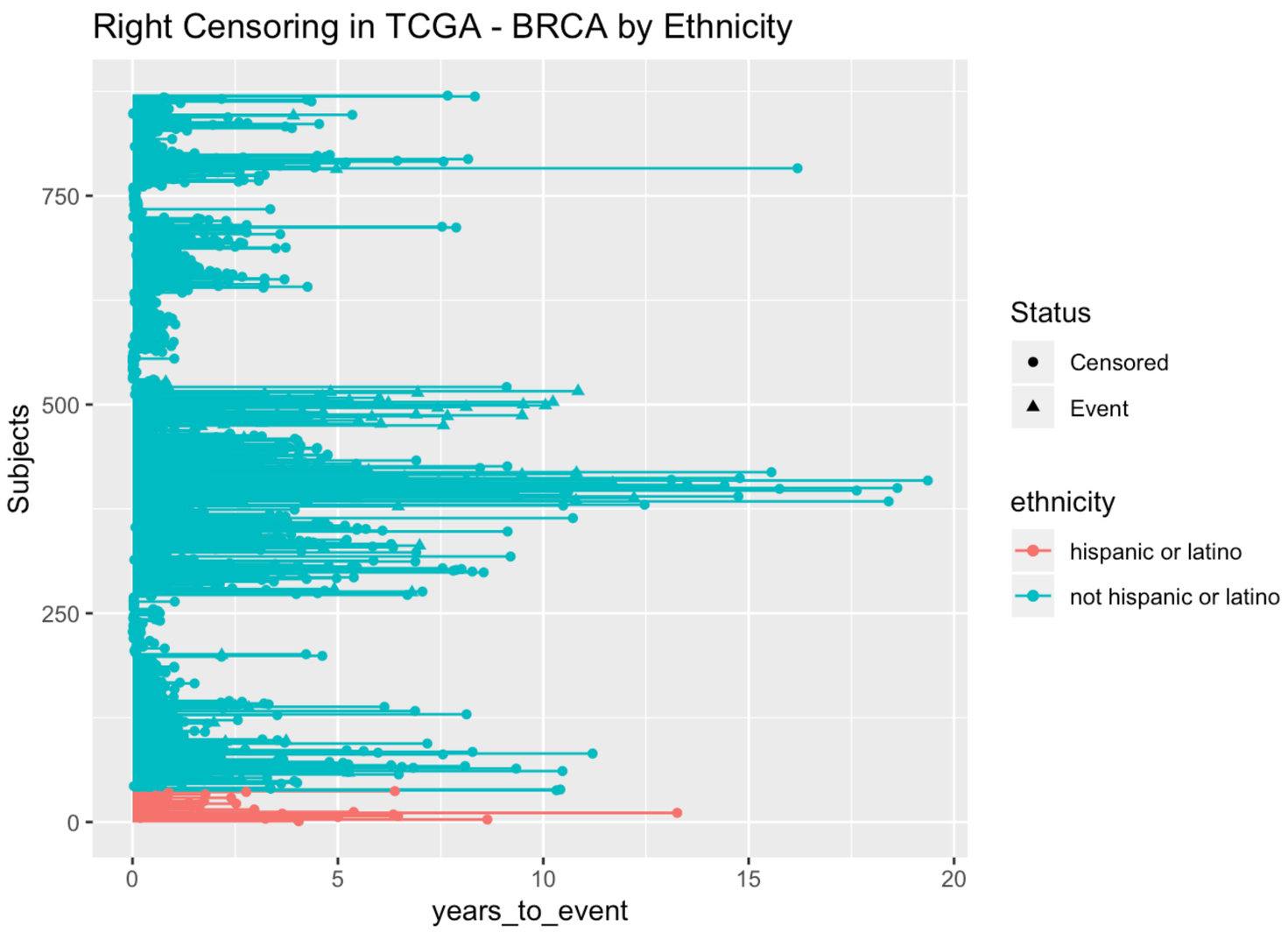
```
##                               vital_status
## race                             0   1 Sum
##   black or african american    160  19 179
##   white                        643  79 722
##   Other                         45   1  46
##   Sum                          848  99 947
```



From these plots, it can be observed that a large percentage of the subjects are white. The number of events is not excessively high in any of the groups.

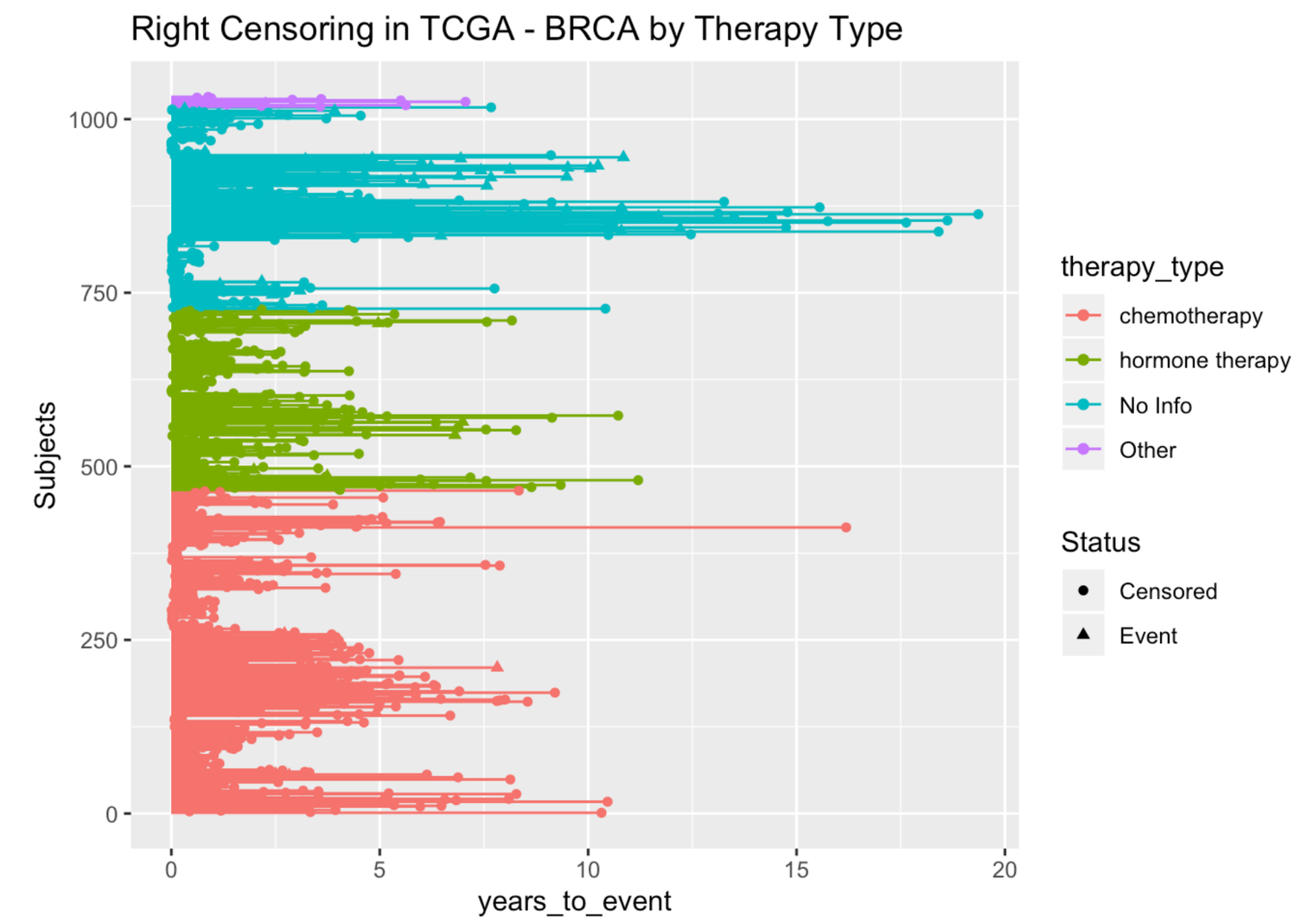
6.4.5 Censoring and Event Plots by Ethnicity

```
##                                vital_status
## ethnicity                         0   1 Sum
##   hispanic or latino             37   0  37
##   not hispanic or latino        737  96 833
##   Sum                           774  96 870
```



6.4.6 Censoring and Event Plots by Therapy Type

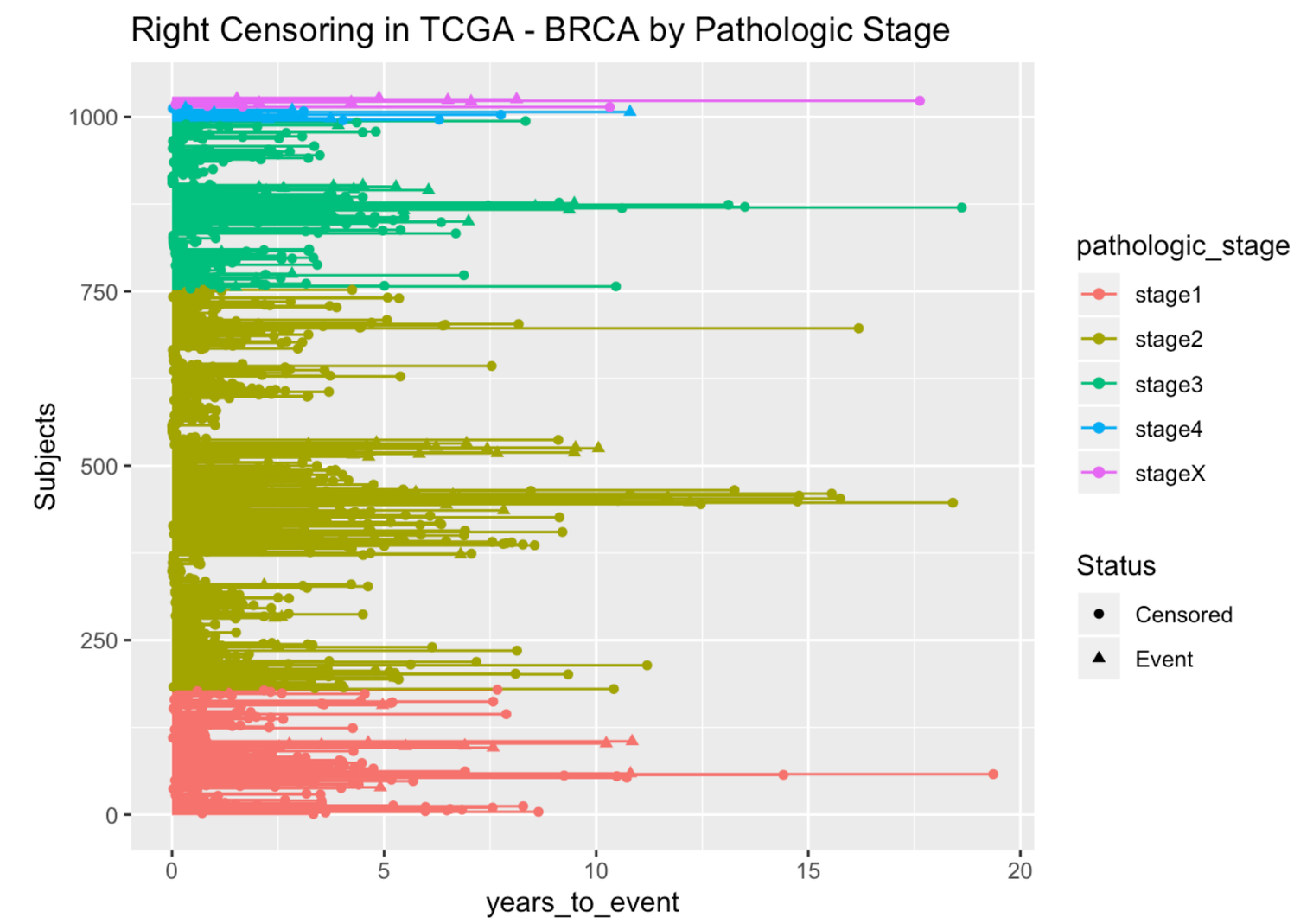
```
##                  vital_status
## therapy_type          0     1   Sum
##   chemotherapy      447    18   465
##   hormone therapy   254     7   261
##   No Info           213    78   291
##   Other              14     1    15
##   Sum               928   104  1032
```



From the contingency table and the plots, it can be observed that there are most of the subjects are on chemotherapy or hormone therapy with more number of people on chemotherapy. The percentage of events is almost the same in both the groups.

6.4.7 Censoring and Event Plots by Pathological Stage

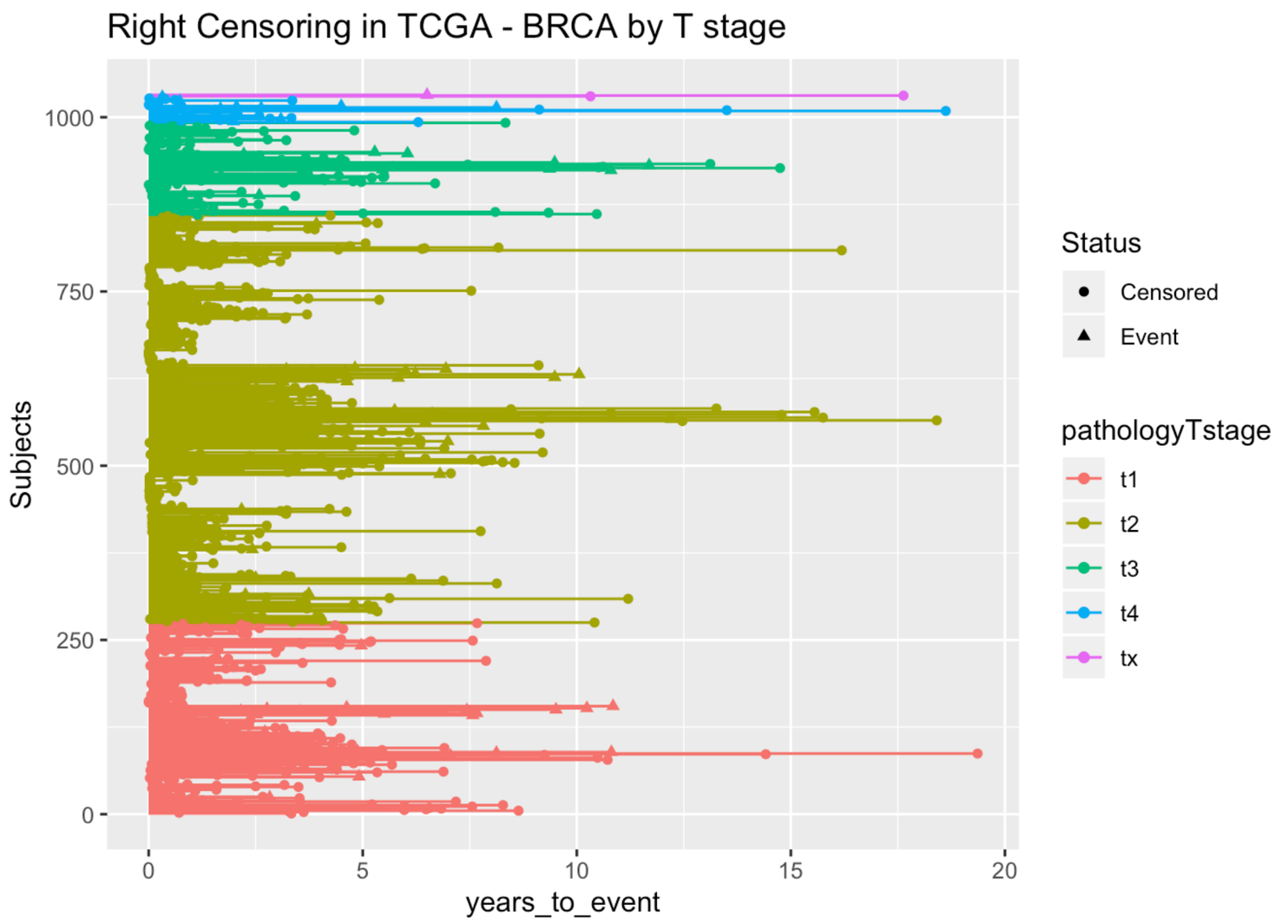
```
##                  vital_status
## pathologic_stage    0     1   Sum
##           stage1  166    13   179
##           stage2  530    44   574
##           stage3  211    30   241
##           stage4   10     9    19
##           stageX    7     7    14
##           Sum     924   103  1027
```



From the contingency table and the censor plots, it can be observed that almost 50% of the subjects are diagnosed with stage 2 cancer, with less than 2% of the subjects are either in stage 4 or in stage X. The percentage of events is the highest (almost 50%) for the stage 4 and stage X groups and the lowest for stage 1.

6.4.8 Censoring and Event Plots by T stage

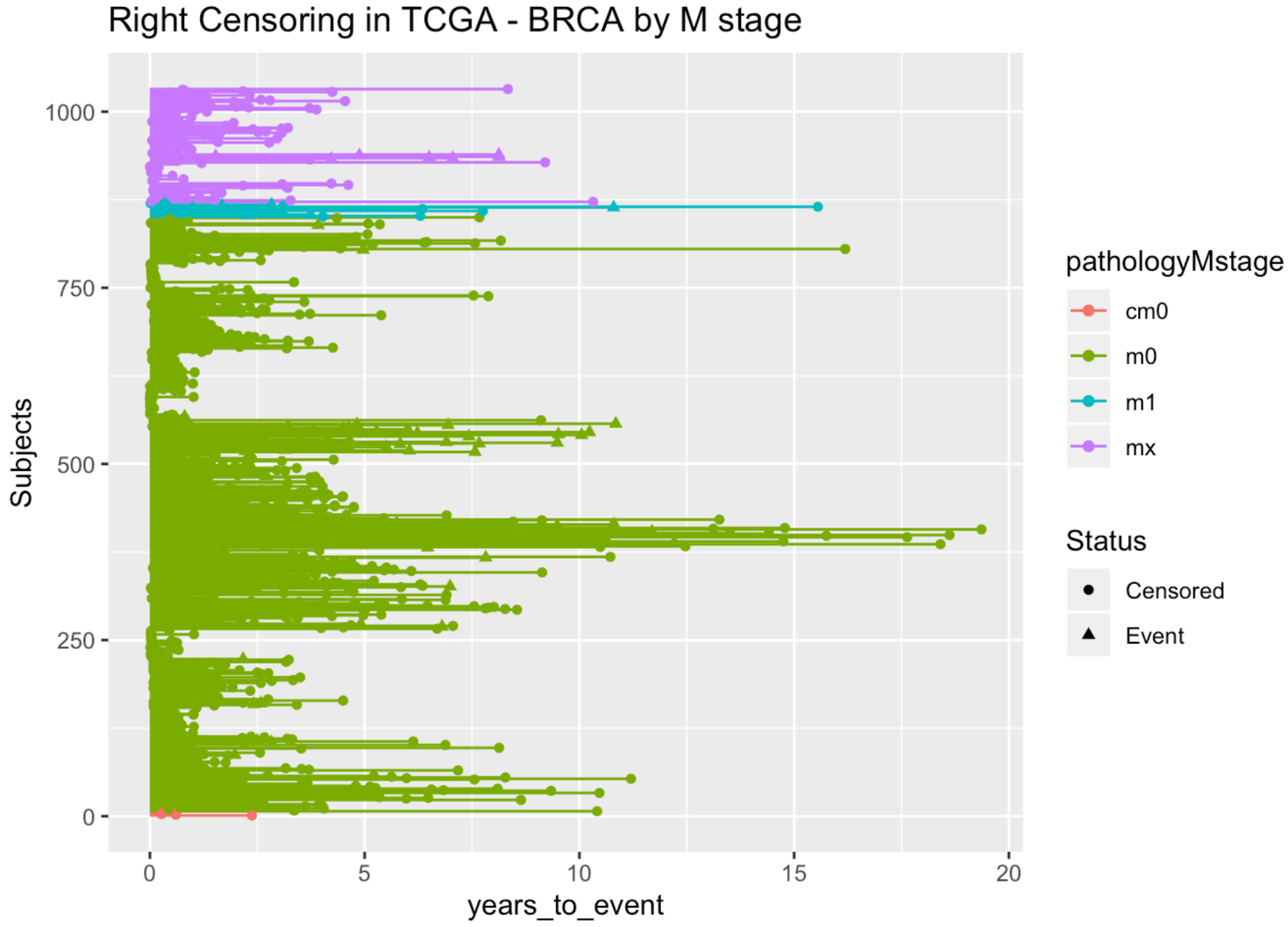
```
##                  vital_status
## pathologyTstage     0     1   Sum
##              t1   248    26   274
##              t2   534    51   585
##              t3   117    16   133
##              t4    27    10    37
##              tx     2     1     3
##              Sum  928   104 1032
```



From the contingency table and the censor plots, it can be observed that the tumors of almost 50% of the subjects are in stage 2, followed by 25% of the subjects with tumors in stge 1. The percentage of events seems to increase with the stage of the tumor.

6.4.9 Censoring and Event Plots by M Stage

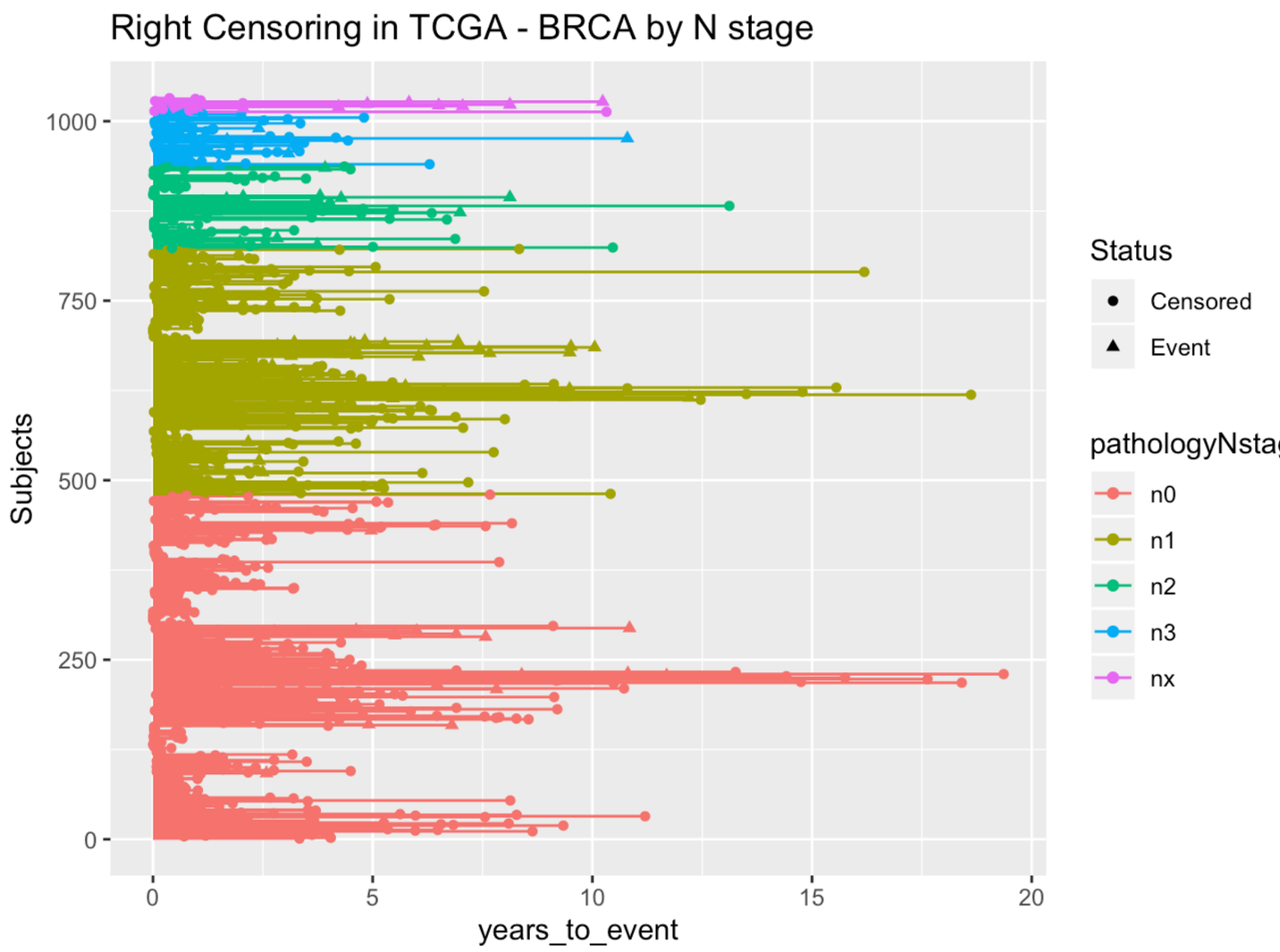
```
##                  vital_status
## pathologyMstage     0     1   Sum
##              cm0     6     0     6
##              m0    757    87   844
##              m1     12     9    21
##              mx    153     8   161
##              Sum   928   104 1032
```



From the contingency table and the plots, it can be observed that in approximately 85% of the subjects, the tumor has not metastised (stage m0 or cm0) and in the few cases where the cancer has metasised (stage m1), the percentage of events is very high.

6.4.10 Censoring and Event Plots by N Stage

	0	1	Sum
n0	452	28	480
n1	300	42	342
n2	101	15	116
n3	64	10	74
nx	11	9	20
Sum	928	104	1032



From the contingency tables and the plots, it can be observed that in almost 50% of the subjects, the cancer has not metasised to the lymph nodes (stage n0). In almost one-third of the subjects, the cancer has metasized to a few lymph nodes (stage n1) and the rest are distriburted between the other groups. The ratio of events to censors increases with the stage number and is highest for stage n4 and stage nx where the status of lymph nodes cannot be assessed.

7 Data Analysis

The final dataset contains 1032 rows and 14 columns. We will do the survival analysis and answer our research questions by modeling the data using the three different methods - non-parametric (Kaplan), semi-parametric (Cox Proportional Hazards), and parametric (Exponential and Weibull). We will start by considering only the time to event data and then repeat the survival analysis by considering the different groups for each of the factors (age category, race, therapy_type, pathology stges - T, M, N). Finally we will include all the variables together for the Cox Regression Analysis.

For each of the predictors (or factors), we will first apply the Kaplan Meirs non-parametric model and plot the hazard and survival curves for each group of the categorical variable and extract the median survival time for each group. We will compare the survival curves between the different levels of each factor and test the hypothesis that the survival curves is the same for each value of the categorical variable. If the curves are found to be significantly different, we will determine the pairs of survival curves that are different.

For each of the predictors, we will also apply the parametric and semi-parametric models to extract more detailed information like hazard ratio which will help us understand the relative risk between the different value of the factors.

7.1 Time to Event only

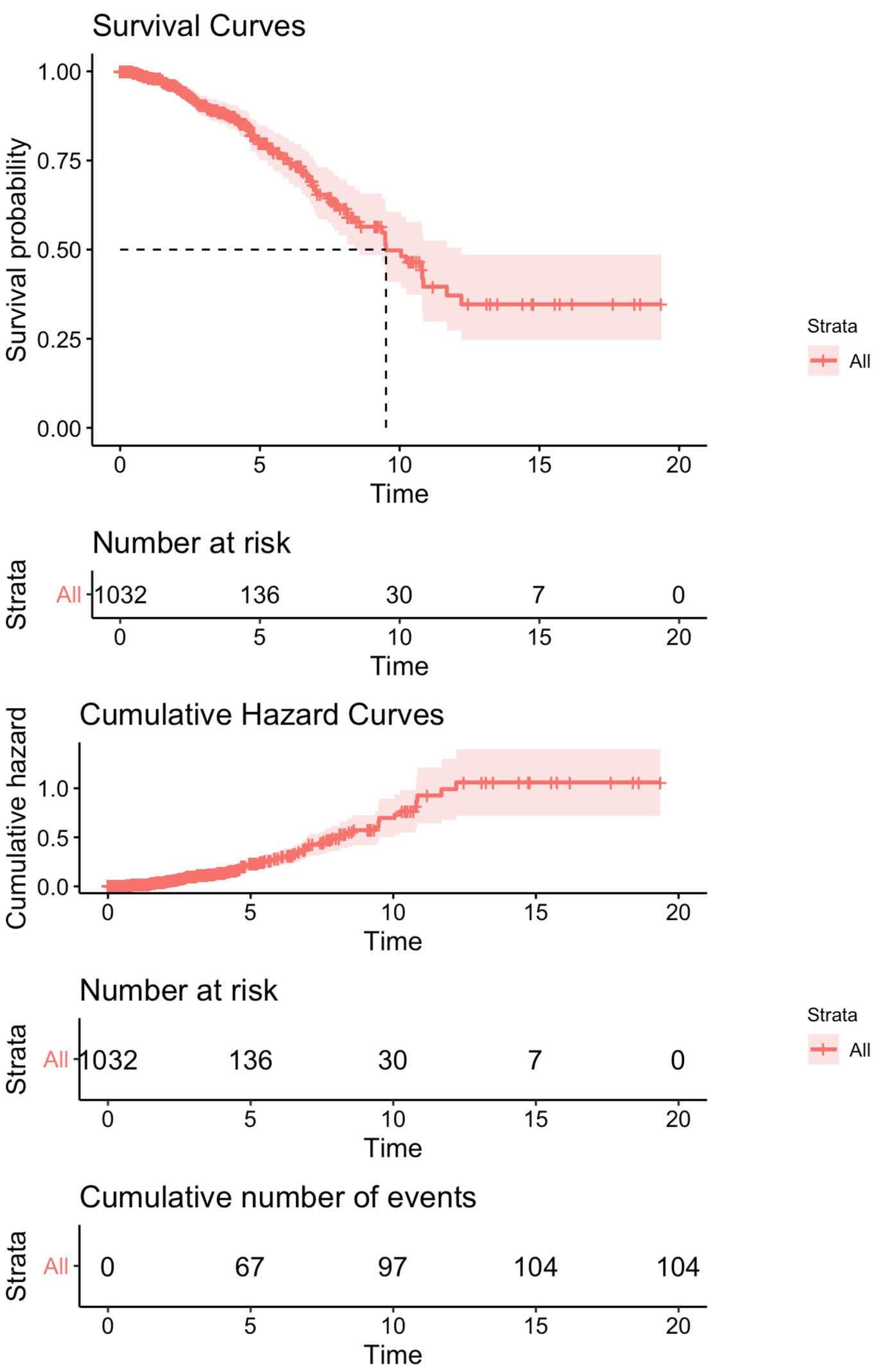
We will first consider only the times data and ignore all other predictors.

7.1.1 Kaplan Meier Survival Curves

```
## Call: survfit(formula = surv_obj ~ 1, data = brca_clin)
##
##       n  events  median 0.95LCL 0.95UCL
## 1032.00  104.00    9.51    8.39   12.21
```

```
## 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.

## 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.
```

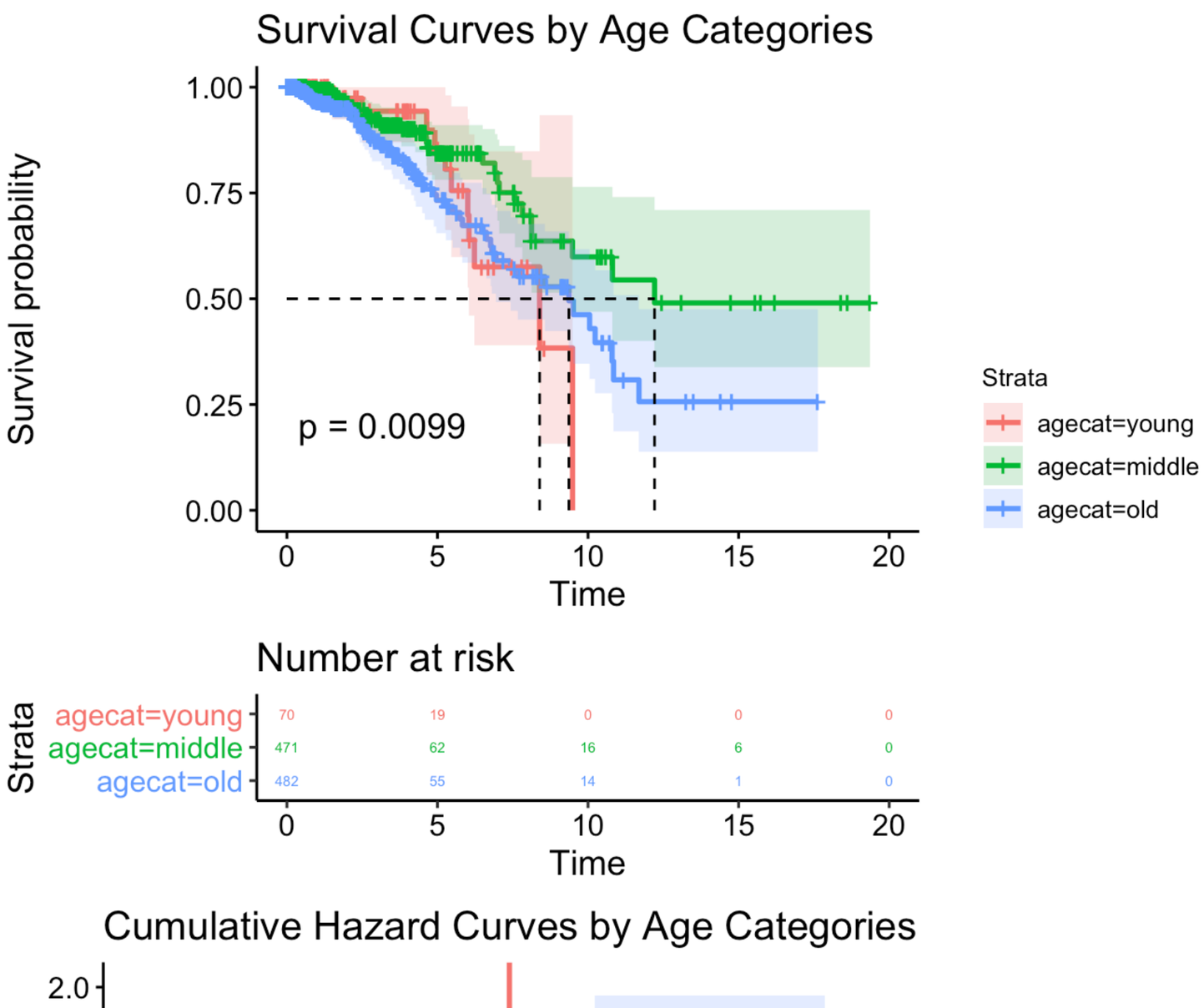


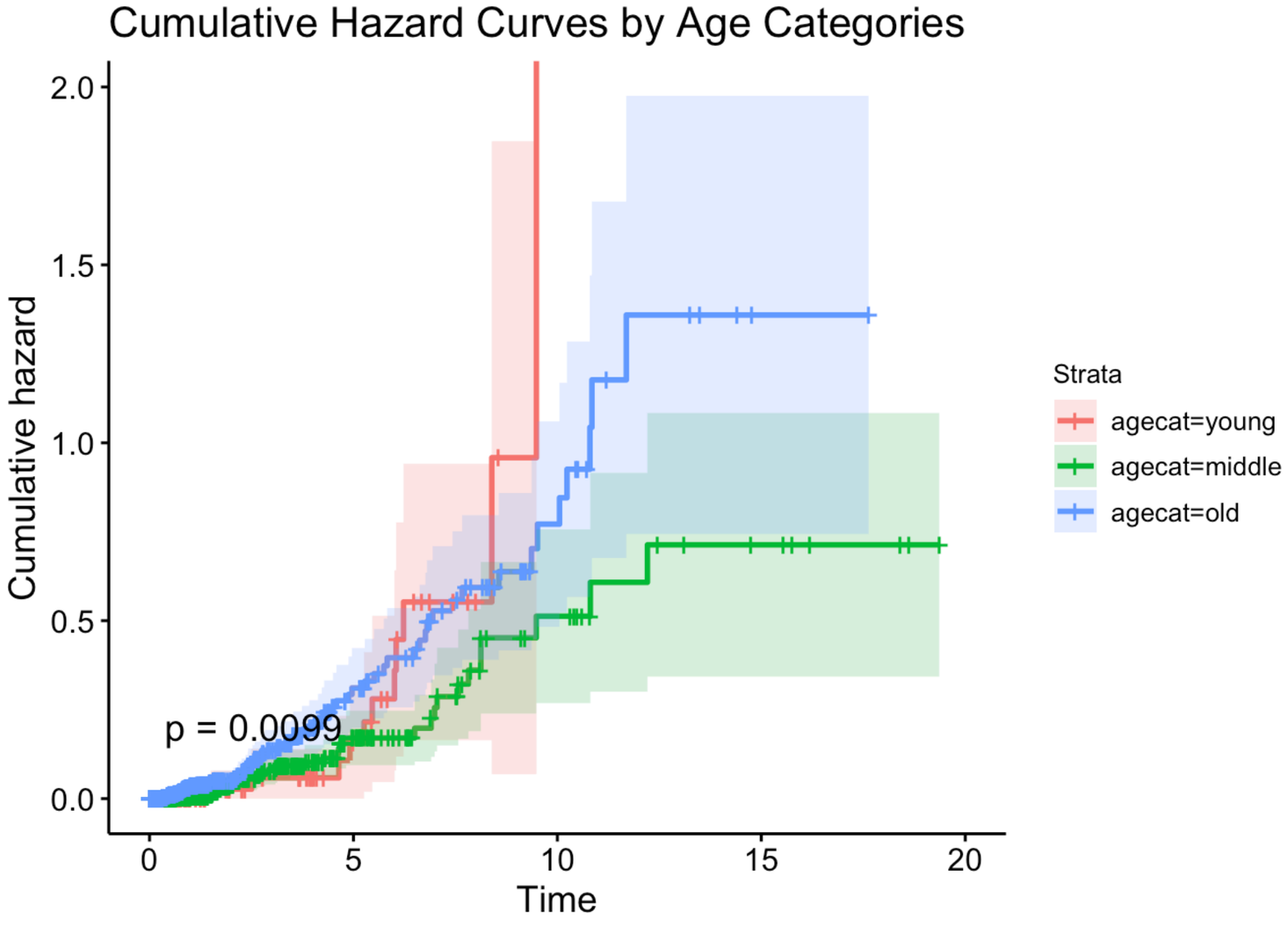
From the survival and cumulative hazard plots, it can be observed that the hazard of death increases very slowly for the first five years, then increases almost linearly from five to ten years and stabilizes around 12 years. The median survival time is 9.5 years.

7.2 Age Category

7.2.1 Kaplan Meier Survival Curves for Age Categories

```
## Call: survfit(formula = surv_obj_age ~ agecat, data = brca_clin_age)
##
##                  n events median 0.95LCL 0.95UCL
## agecat=young    70     11   8.39    6.05      NA
## agecat=middle 471     35  12.21    9.48      NA
## agecat=old    482     58   9.36    6.94    11.7
```





From the survival curves and the cumulative hazard curves plot, it can be observed that the hazard rate for the middle and the old age groups are very similar up to five years and then the hazard rate increases at a faster rate for the older group from 5 to 12 years when it finally becomes steady and remains constant after 12 years for both the groups.

In contrast, the hazard rate increases very rapidly during the 5 to 10 years after diagnosis for the young group. The median survival time is the highest for the middle age group at 12.21 years and at 9.36 for the old age group followed by 8.4 years for the young age group. We have 95% confidence that the survival time for the older age group varies between 6.94 and 11.7 years and there is no upper confidence limit for the young and middle age groups.

Compare Survival Curves

```
## Call:
## survdiff(formula = surv_obj_age ~ agecat, data = brca_clin_age)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## agecat=young    70       11     9.23     0.341     0.381
## agecat=middle 471       35    50.35     4.680     9.133
## agecat=old    482       58    44.42     4.150     7.285
##
##  Chisq= 9.2  on 2 degrees of freedom, p= 0.01
```

When the survival curves for the three groups are compared, it is found that the log-rank statistic or the MH statistic with a chi square distribution has a large value of 9.2 and a p-vaue of 0.01. Using a confidence limit of 0.05, we have sufficient statistical evidence to reject the null hypothesis and conclude that there is significant difference in the survival curves for the different groups.

Check for pairs of curves that are different

```
##
##  Pairwise comparisons using Log-Rank test
##
## data:  brca_clin and agecat
##
##        young  middle
## middle 0.1949 -
## old    0.8015 0.0069
##
## P value adjustment method: BH
```

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

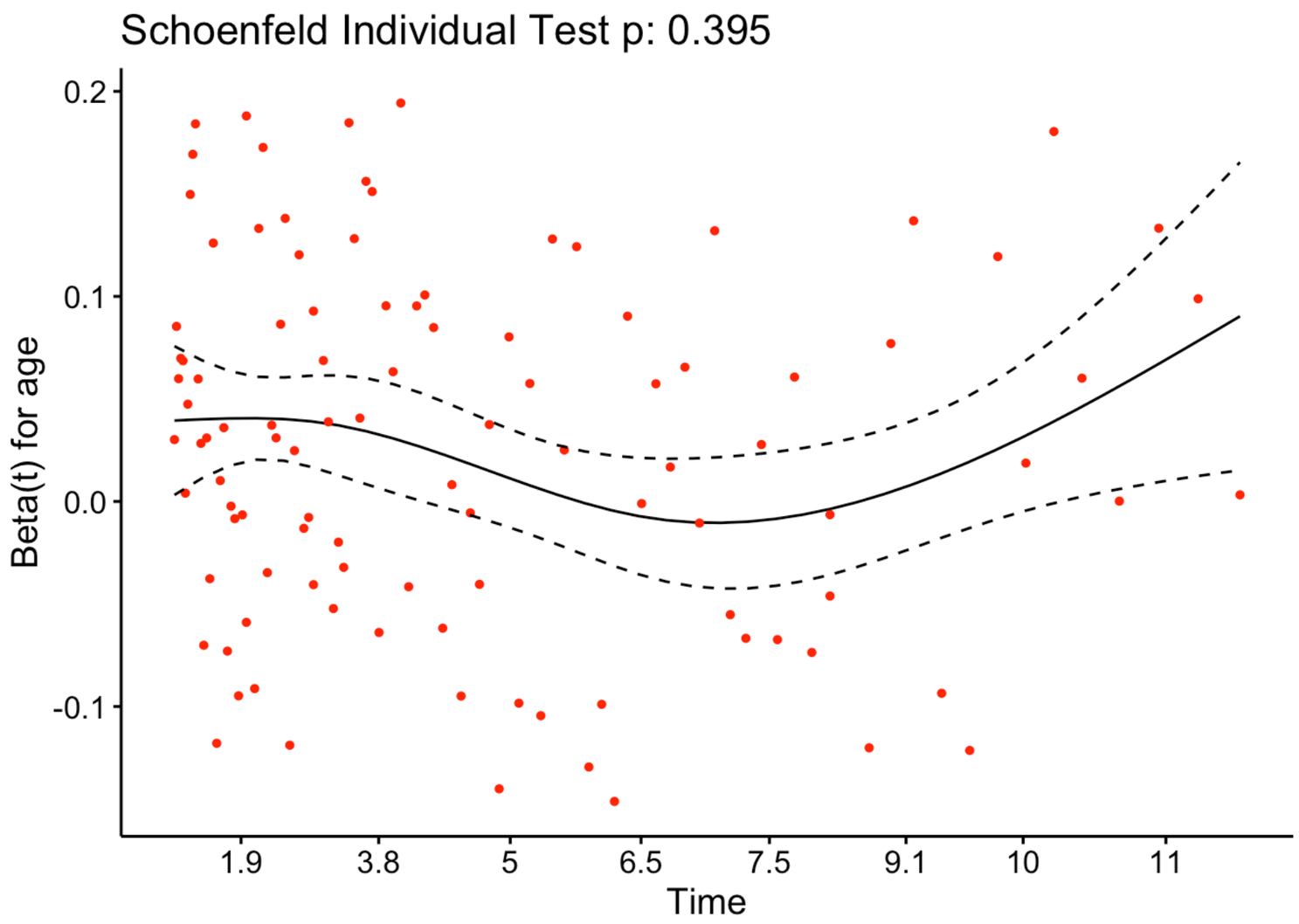
On doing a pairwise comparison of the curves, the survival curves of the middle and old age groups are significantly different.

7.2.2 Cox Proportional Hazard (PH) Model for Age

The Cox PH model is fit using age as the predictor. To do this, we use use as a continuous variable instead of as a categorical variable as was used with Kaplan Meiers model.

Check for Proportional Hazard Assumptions

```
##          rho chisq     p
## age -0.0721 0.724 0.395
```



The Schonfeld residuals and the p-value for the corresponding chi-square distribution are checked to verify the assumption of proportional hazards of the Cox model. The plot of the Schonfeld residuals shows a random distibution around a mean of 0. The corresponding p-value for the chi-square distribution is high indicating non-significance. This indicates that the proportional hazards condition is met.

Interpret Model Output

```
## Call:
## coxph(formula = surv_obj_age ~ age, data = brca_clin_age)
##
##   n= 1023, number of events= 104
##
##         coef exp(coef) se(coef)    z Pr(>|z|)
## age 0.026513  1.026868 0.007407 3.58 0.000344 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## age     1.027     0.9738     1.012     1.042
##
## Concordance= 0.63  (se = 0.034 )
## Likelihood ratio test= 12.77  on 1 df,   p=4e-04
## Wald test            = 12.81  on 1 df,   p=3e-04
## Score (logrank) test = 12.99  on 1 df,   p=3e-04
```

The p-value of the Wald test for age is less than 0.05 . This indicates that we can reject the null hypothesis that the coefficient for age is zero and conclude that the age is indeed a significant factor in determining the hazard rate for a person with breast cancer. The Cox model equation can then be written as:

$$H(t|age) = H_0(t)exp(0.0265 * age)$$

This equation can be used to compute the cumulative hazard rate for any age.

From the Cox PH model, the hazard ratio (exp(coef)) is 1.027 indicating the hazard rate increases by a factor of 1.027 for every increase in age by 1 year.

7.2.3 Parametric Modeling - Age as Catergorical

```
##
## Call:
## survreg(formula = surv_obj_age ~ agecat, data = brca_clin_age,
##     dist = "weibull")
##                Value Std. Error     z       p
## (Intercept)   2.4435     0.1929 12.67 < 2e-16
## agecatmiddle  0.3766     0.2149  1.75    0.08
## agecatold    -0.0482     0.2043 -0.24    0.81
## Log(scale)   -0.4772     0.0642 -7.43 1.1e-13
##
## Scale= 0.621
##
## Weibull distribution
## Loglik(model)= -404.4   Loglik(intercept only)= -410
##  Chisq= 11.12 on 2 degrees of freedom, p= 0.0038
## Number of Newton-Raphson Iterations: 17
## n= 1023
```

```
##
## Call:
## survreg(formula = surv_obj_age ~ agecat, data = brca_clin_age,
##     dist = "exponential")
##                Value Std. Error     z      p
## (Intercept)   2.9452     0.3015  9.77 <2e-16
## agecatmiddle  0.5675     0.3457  1.64   0.10
## agecatold    -0.0683     0.3289 -0.21   0.84
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -426.2   Loglik(intercept only)= -431
##  Chisq= 9.6 on 2 degrees of freedom, p= 0.0082
## Number of Newton-Raphson Iterations: 7
## n= 1023
```

The survival data is fit with parametric models using both the exponential and weibull distributions and compared. The fitness of the each of these models is verified by checking the loglikelihood values of the models. For both the models, p-values for the chi-square distribution with one degree of freedom are very small (<0.05) indicating that models are good fits.

```
## [1] "The better fit parametric model: weibull"
```

Anova is used to compare the two fitted models and it is found that the two model fits are significantly different with the weibull distribution model being the better fit with the lower loglikelihood value.

```
## [[1]]
##                  Estimate          SE
## lambda         0.01949249 0.006818777
## gamma          1.61150274 0.103445374
## agecatmiddle -0.60696042 0.346074988
## agecatold      0.07767801 0.328986500
```

From the converted results of the weibull model, the risk of death for the middle age group (ages 40-60) is almost 45% less (HR=0.55) than the risk of death for the younger group (age < 40) and the risk of death for the older age group (age > 60) is slightly higher by 10% compared to the younger age group.

Equivalently, the survival time for the middle age group is higher by 46 % (ETR=1.46) for the middle age group when compared with the younger group and the survival time for the older age group is less by 95% when compared with the younger age group.

7.2.4 Parameter Modeling Age as continuous

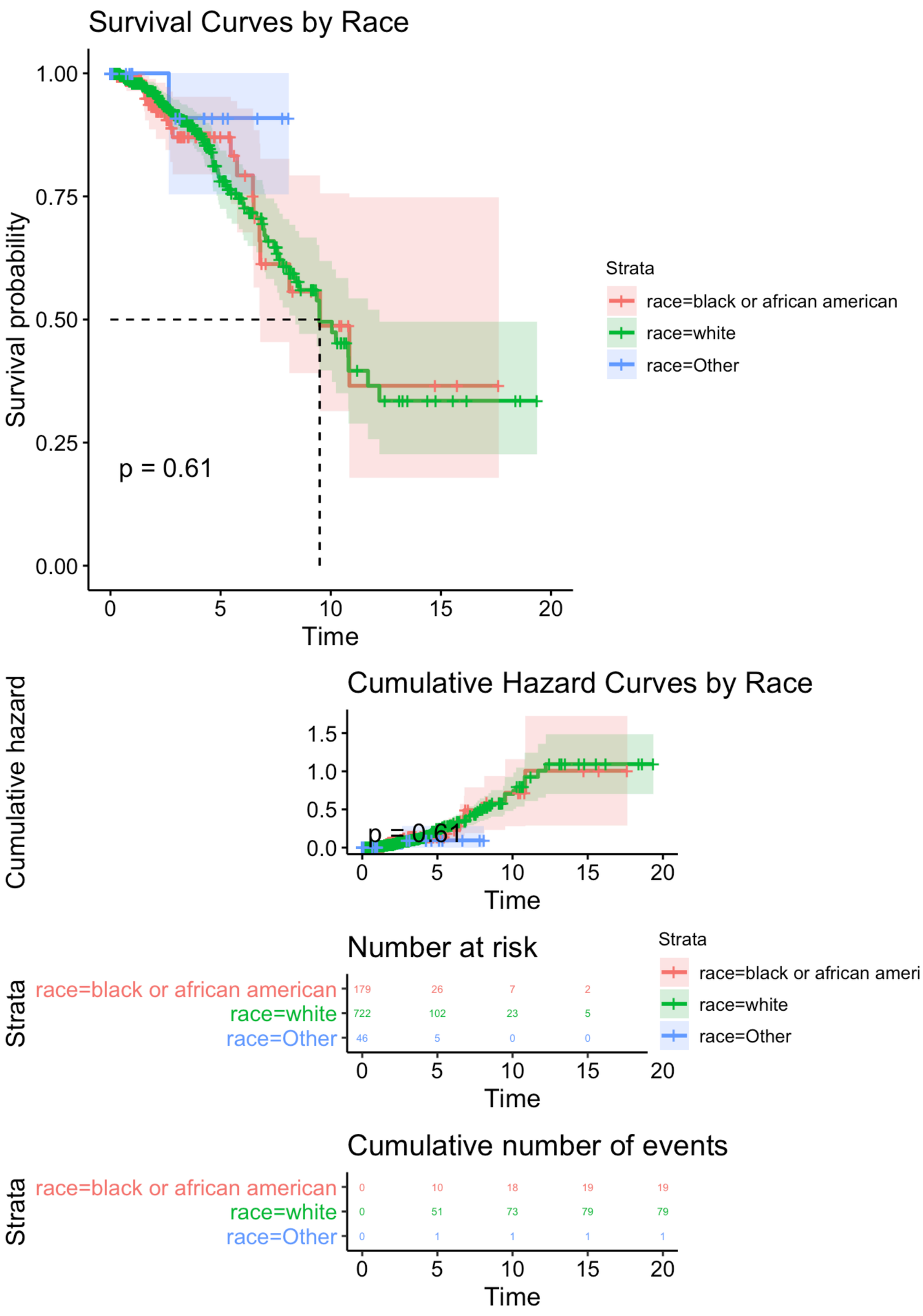
```
##
## Call:
## survreg(formula = surv_obj_age ~ age, data = brca_clin_age, dist = "weibull")
##                 Value Std. Error     z       p
## (Intercept)  3.60754    0.29916 12.06 < 2e-16
## age         -0.01761    0.00459 -3.84 0.00013
## Log(scale)  -0.48644    0.06366 -7.64 2.2e-14
##
## Scale= 0.615
##
## Weibull distribution
## Loglik(model)= -402.6   Loglik(intercept only)= -410
##  Chisq= 14.9 on 1 degrees of freedom, p= 0.00011
## Number of Newton-Raphson Iterations: 13
## n= 1023
```

```
##
## Call:
## survreg(formula = surv_obj_age ~ agecat, data = brca_clin_age,
##     dist = "exponential")
##                Value Std. Error     z      p
## (Intercept)   2.9452     0.3015  9.77 <2e-16
## agecatmiddle  0.5675     0.3457  1.64   0.10
## agecatold    -0.0683     0.3289 -0.21   0.84
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -426.2   Loglik(intercept only)= -431
##  Chisq= 9.6 on 2 degrees of freedom, p= 0.0082
## Number of Newton-Raphson Iterations: 7
## n= 1023
```

7.3 Race

7.3.1 Kaplan Meier Survival Curves

```
## Call: survfit(formula = surv_obj_race ~ race, data = brca_clin_race)
##
##                                   n events median 0.95LCL 0.95UCL
## race=black or african american 179     19   9.51    6.80      NA
## race=white                     722     79   9.48    8.39    12.2
## race=Other                      46      1     NA      NA      NA
```




```
## Call:
## survdiff(formula = surv_obj_race ~ race, data = brca_clin_race)
##
##                                 N Observed Expected (O-E)^2/E (O-E)^2/V
## race=black or african american 179       19    18.61   0.00817    0.0101
## race=white                     722       79    77.82   0.01790    0.0838
## race=Other                      46        1     2.57   0.95908    0.9903
##
##  Chisq= 1  on 2 degrees of freedom, p= 0.6
```

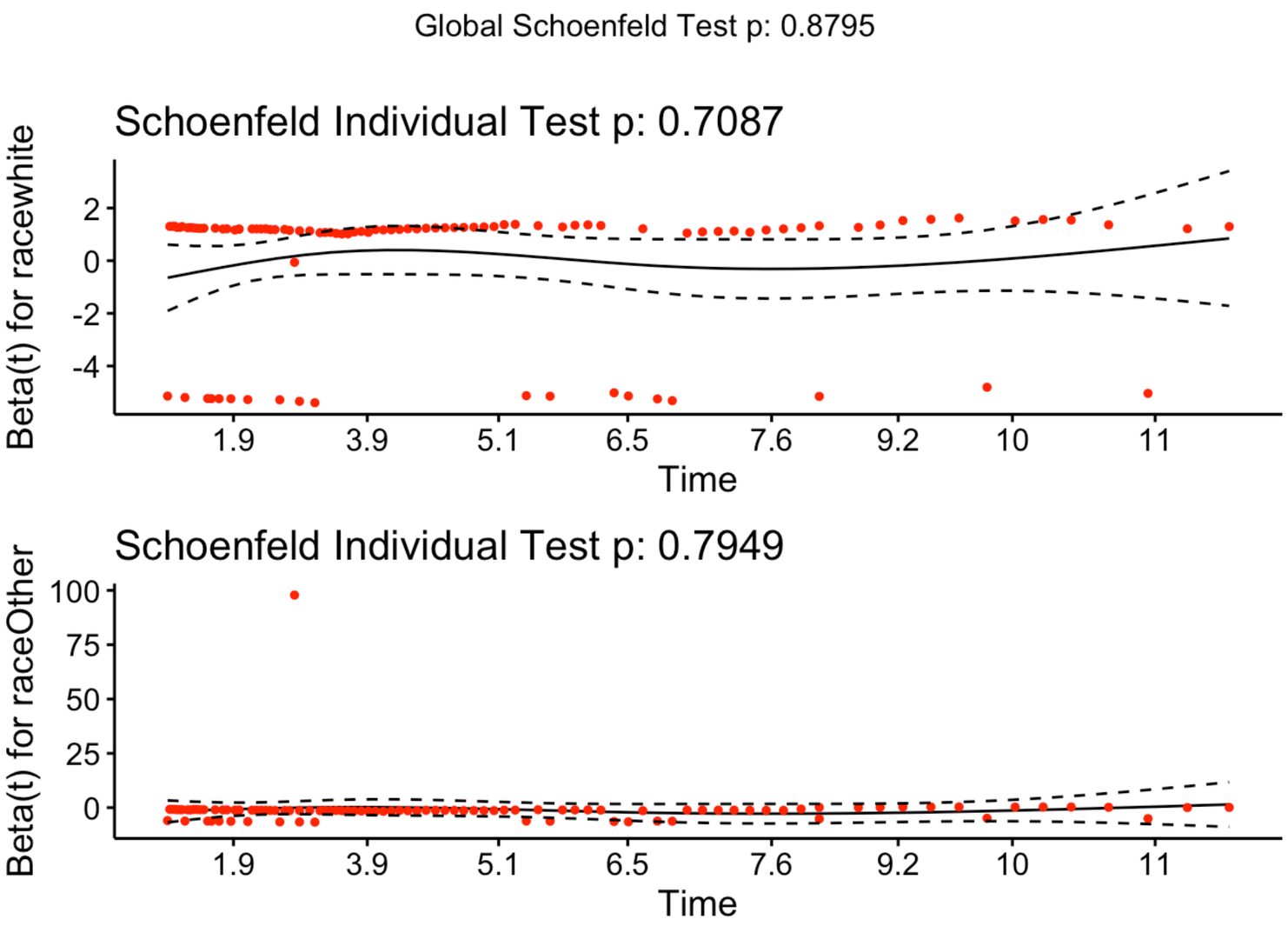
From the survival curves plot, it can be observed that the survival curves for both the races - white and black or african-american are very similar. The probablity of survival steadily reduces during the first 12 years and then remains flat. The median survival time for the both the races is the same at 9.5 years. This means that 50% of the people are expected to survive after 9.5 years since first diagnosis.

From the output of survdiff, it can be seen that the log-rank statistic or the MH statistic with a chi square distribution has a value close to 1 and a p-vaue of 0.8. Using a confidence limit of 0.05, we do not have sufficient evidence to reject the null hypothesis and conclude that there is no difference in survival rates for the different groups.

7.3.2 Cox Proportional Hazard Model for Race

Check for Proportional Hazard Assumptions

```
##                rho  chisq     p
## racewhite   0.0378 0.1395 0.709
## raceOther -0.0262 0.0676 0.795
## GLOBAL          NA 0.2569 0.879
```

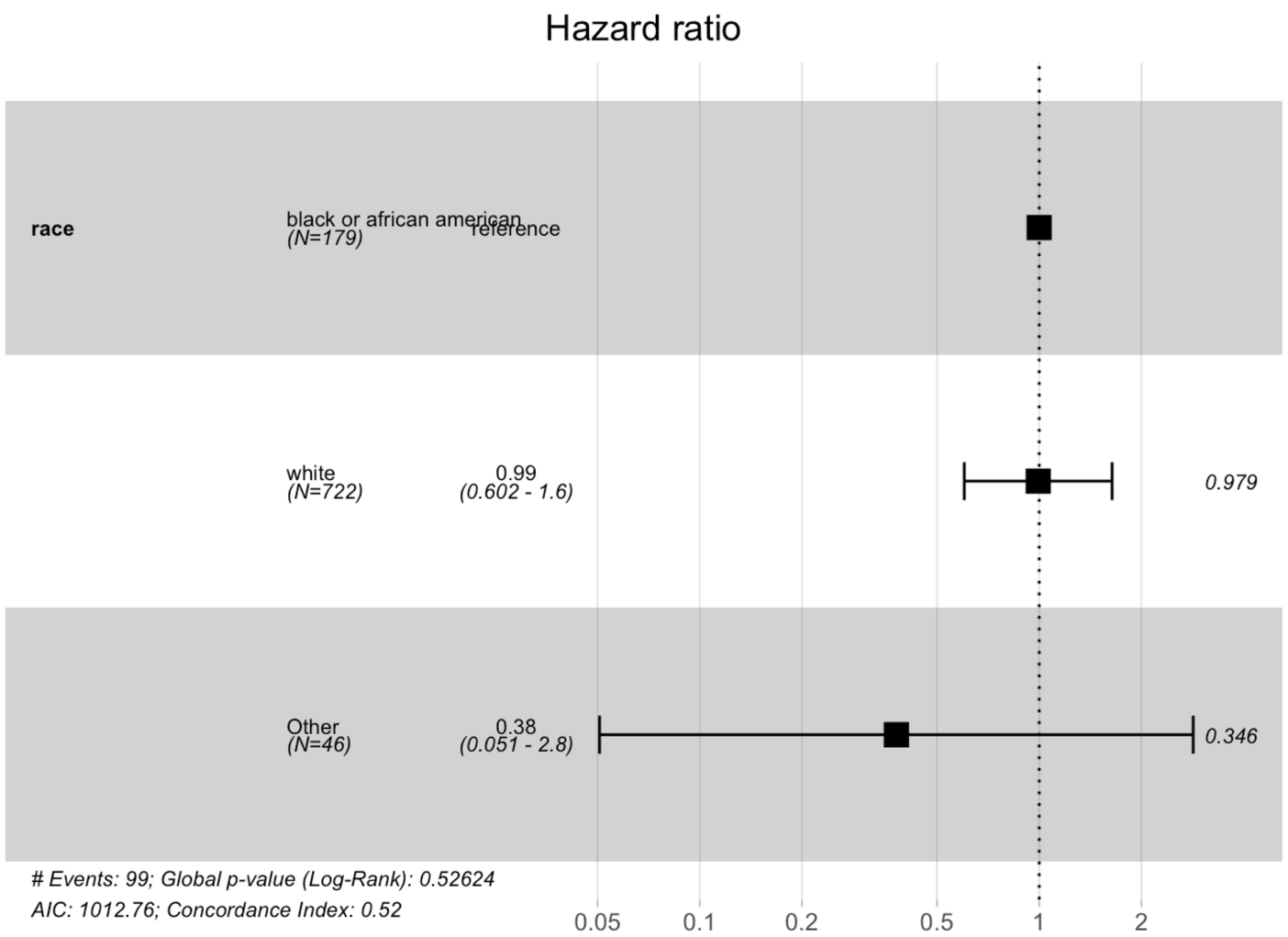


The proportional hazards assumption of the Cox model is verified by checking the Schonfeld residuals. The plot of the Schonfeld residuals shows a random distibution around a mean of 0. The corresponding p-values for the chi-square distribution are high indicating non-significance. This indicates that the proportional hazards condition is met.

Interpret the model

```
## Call:
## coxph(formula = surv_obj_race ~ race, data = brca_clin_race)
##
##   n= 947, number of events= 99
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## racewhite -0.006724  0.993298  0.255838 -0.026    0.979
## raceOther -0.969064  0.379438  1.027398 -0.943    0.346
##
##           exp(coef) exp(-coef) lower .95 upper .95
## racewhite    0.9933      1.007   0.60160     1.640
## raceOther    0.3794      2.635   0.05065     2.842
##
## Concordance= 0.517  (se = 0.027 )
## Likelihood ratio test= 1.28  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
```

```
## Warning: Removed 1 rows containing missing values (geom_errorbar).
```



The p-values of the Wald test for both the races (white and Other) are large indicating that the coefficients for both these predictors are not significantly different from zero. This means that hazards are the same for each race group. This is also evident from the confidence intervals of the hazard ratio as they include the 0 value. So, the Cox model does not show significant difference in hazards. between the different race groups.

7.3.3 Parametric modeling - Race

```
## Warning in survreg.fit(X, Y, weights, offset, init = init, controlvals =
## control, : Ran out of iterations and did not converge
```

```
## $vars
##               Estimate          SE
## lambda     0.005544232 0.001270994
## gamma      2.126526013 0.000000000
## racewhite  0.022562878 0.255372435
## raceOther  6.724972909 0.230446394
##
## $HR
##                    HR          LB          UB
## racewhite    1.022819   0.6200483    1.687222
## raceOther  832.949411 530.2275620 1308.503689
##
## $ETR
##                   ETR         LB         UB
## racewhite  0.98944588 0.78193757 1.25202214
## raceOther  0.04232311 0.03422433 0.05233836
```

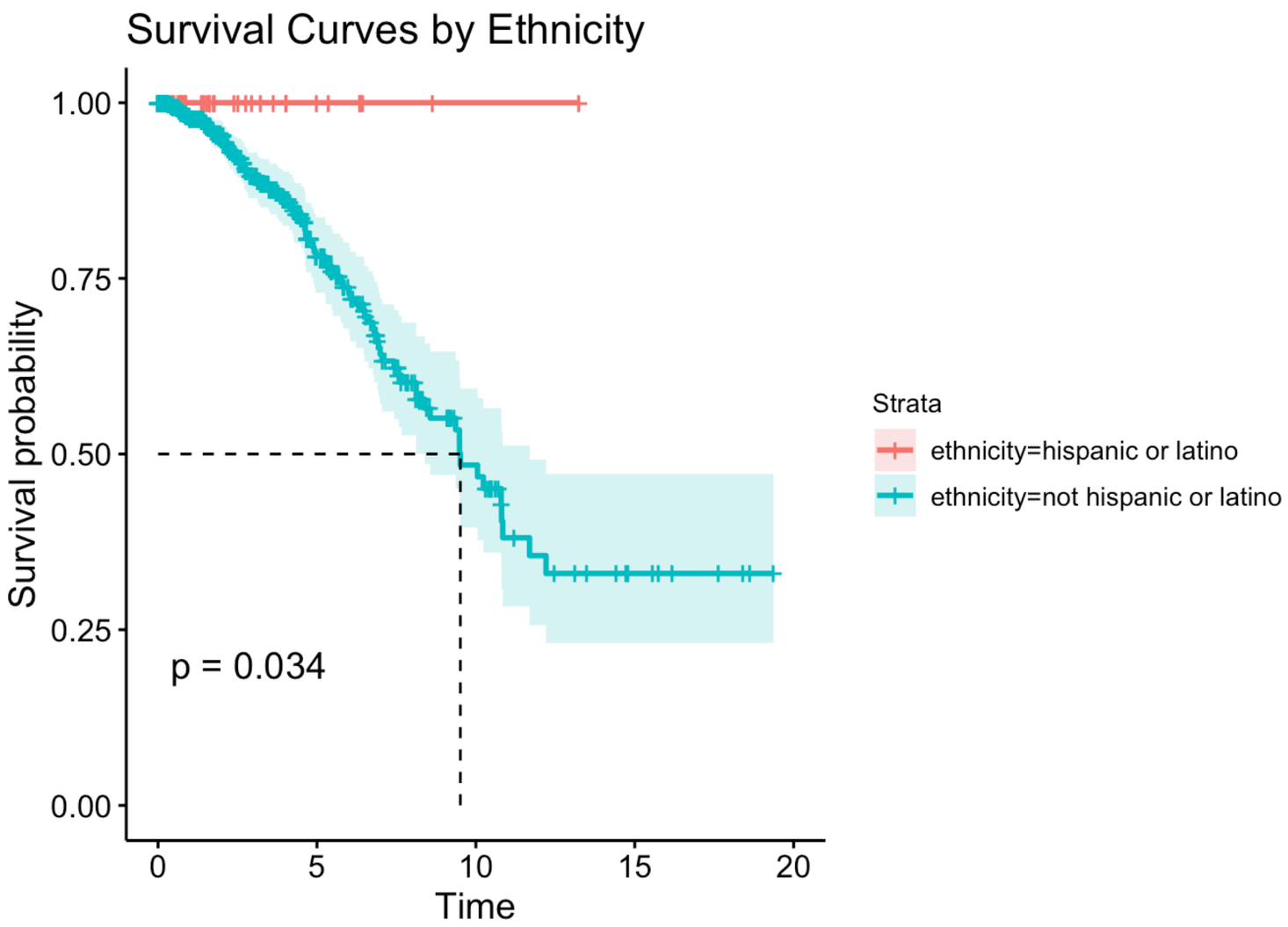
```
##
## Call:
## survreg(formula = surv_obj_race ~ race, data = brca_clin_race,
##     dist = "exponential")
##              Value Std. Error     z      p
## (Intercept) 3.1155     0.2294 13.58 <2e-16
## racewhite   0.0117     0.2555  0.05   0.96
## raceOther   0.9836     1.0260  0.96   0.34
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -409.3   Loglik(intercept only)= -410
##  Chisq= 1.32 on 2 degrees of freedom, p= 0.52
## Number of Newton-Raphson Iterations: 9
## n= 947
```

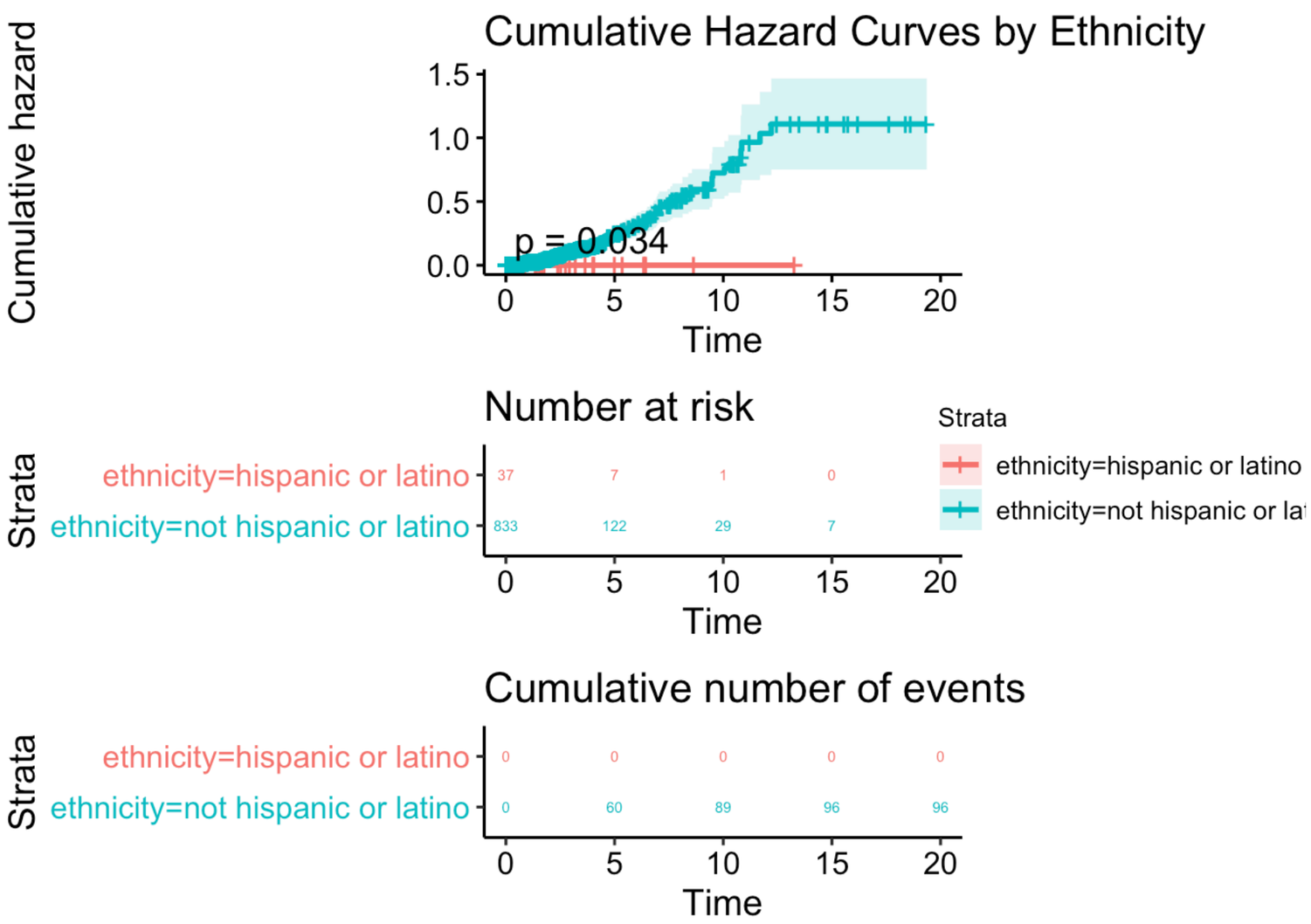
The survival data is fit with parametric models using both the exponential and weibull distributions and compared. The fitness of the each of these models is verified by checking the loglikelihood values of the models. For both the models, p-values for the chi-square distribution with one degree of freedom are high (> 0.05) indicating that models are not good fits.

7.4 Ethnicity

7.4.1 Kaplan Meier Survival Curves for Ethnicity

```
## Call: survfit(formula = surv_obj_ethnicity ~ ethnicity, data = brca_clin_ethnicity)
##
##                                     n events median 0.95LCL 0.95UCL
## ethnicity=hispanic or latino       37      0     NA      NA      NA
## ethnicity=not hispanic or latino 833     96   9.51    8.39    11.7
```





The groups are very unbalanced with very few subjects in the non-hispanic group. So, the results of the survival curves being very different may not be very meaningful.

7.4.2 Cox Proportional Hazard (PH) Model for Ethnicity

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

The Cox model is not a good fit for this predictor as infinite values get introduced. This may be due to the balanced groups contributing to complete separation with none of the subjects in the hispanic group experiencing an event.

7.4.3 Parametric Modeling - Ethnicity

```
##
## Call:
## survreg(formula = surv_obj_ethnicity ~ ethnicity, data = brca_clin_ethnicity,
##     dist = "weibull")
##                                   Value Std. Error     z       p
## (Intercept)                     14.0531  2323.1464  0.01       1
## ethnicitynot hispanic or latino -11.4893  2323.1464  0.00       1
## Log(scale)                      -0.4409     0.0679 -6.49 8.4e-11
##
## Scale= 0.643
##
## Weibull distribution
## Loglik(model)= -370.6   Loglik(intercept only)= -374.9
##  Chisq= 8.52 on 1 degrees of freedom, p= 0.0035
## Number of Newton-Raphson Iterations: 26
## n= 870
```

```
##
## Call:
## survreg(formula = surv_obj_ethnicity ~ ethnicity, data = brca_clin_ethnicity,
##     dist = "exponential")
##                                 Value Std. Error     z    p
## (Intercept)                      20.4     2780.1  0.01 0.99
## ethnicitynot hispanic or latino -17.4     2780.1 -0.01 1.00
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -387.7   Loglik(intercept only)= -392.2
##  Chisq= 9.06 on 1 degrees of freedom, p= 0.0026
## Number of Newton-Raphson Iterations: 19
## n= 870
```

The survival data is fit with parametric models using both the exponential and weibull distributions and compared. The fitness of the each of these models is verified by checking the loglikelihood values of the models. For both the models, p-values for the chi-square distribution with one degree of freedom are very small (<0.05) indicating that models are good fits.

```
## [1] "The better fit parametric model: weibull"
```

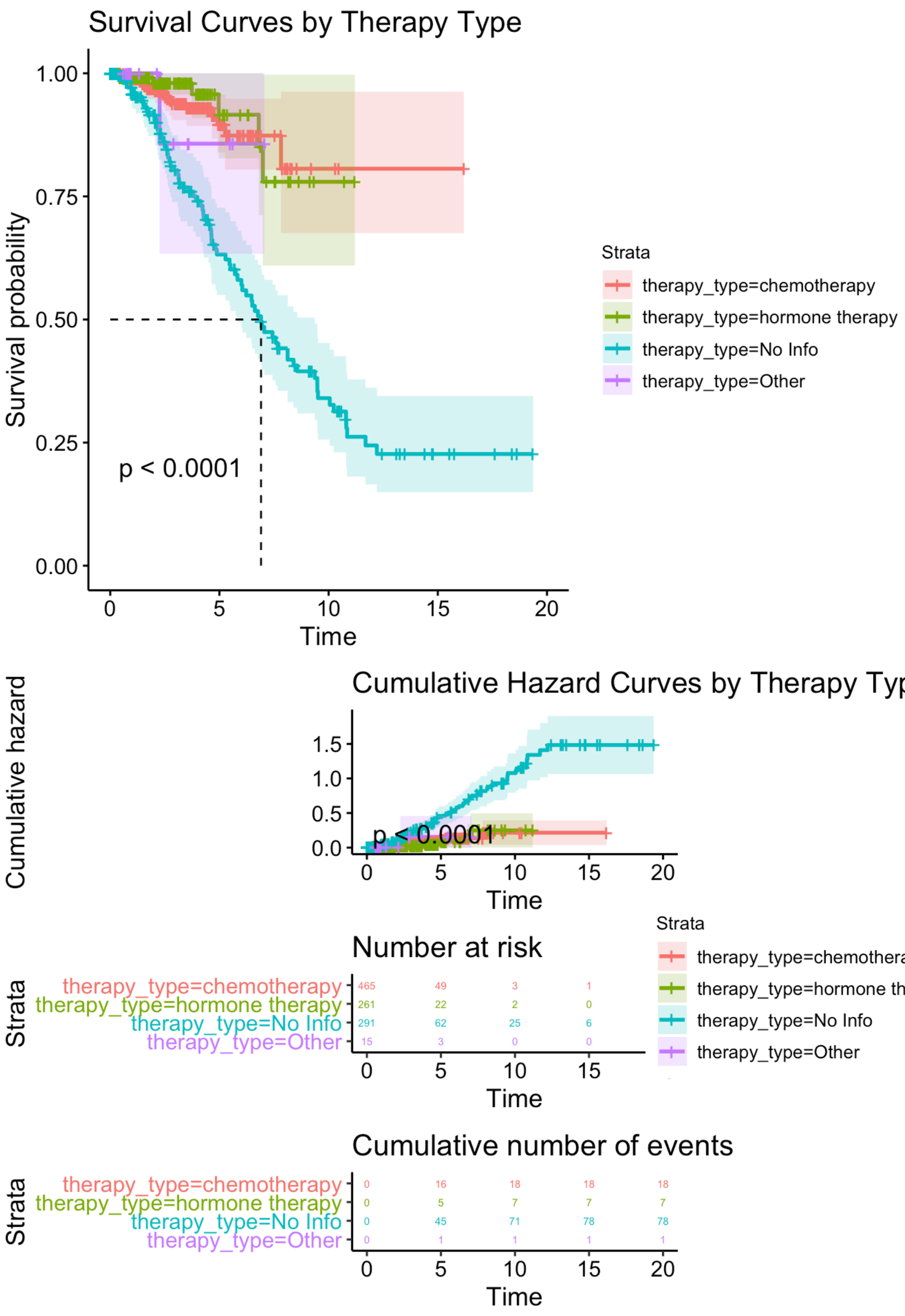
Anova is used to compare the two fitted models and it is found that the two model fits are significantly different with the weibull distribution model being the better fit with the lower loglikelihood value.

When the converted weibull values are considered, it is found the hazard rates cannot be modeled correctly resulting in infinite values.

7.5 Therapy Type

7.5.1 Kaplan Meier Survival Curves for Therapy Type

```
## Call: survfit(formula = surv_obj_therapy ~ therapy_type, data = brca_clin_therapy)
##
##                                n events median 0.95LCL 0.95UCL
## therapy_type=chemotherapy    465     18     NA      NA      NA
## therapy_type=hormone therapy 261      7     NA      NA      NA
## therapy_type=No Info         291     78    6.9    5.83    9.36
## therapy_type=Other            15      1     NA      NA      NA
```




```
## Call:
## survdiff(formula = surv_obj_therapy ~ factor(therapy_type, exclude = NULL),
##     data = brca_clin_therapy)
##
##                                                  N Observed Expected
## factor(therapy_type, exclude = NULL)=chemotherapy    465       18     36.3
## factor(therapy_type, exclude = NULL)=hormone therapy 261        7     19.0
## factor(therapy_type, exclude = NULL)=No Info         291       78     47.2
## factor(therapy_type, exclude = NULL)=Other            15        1      1.5
##                                                 (O-E)^2/E (O-E)^2/V
## factor(therapy_type, exclude = NULL)=chemotherapy     9.247    15.458
## factor(therapy_type, exclude = NULL)=hormone therapy  7.543     9.467
## factor(therapy_type, exclude = NULL)=No Info         20.083    43.810
## factor(therapy_type, exclude = NULL)=Other            0.169     0.172
##
##  Chisq= 44  on 3 degrees of freedom, p= 1e-09
```

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

From the plots of the survival curves and the cumulative hazard rate, it can be observed that there is a big difference in survival and hazard rates between the groups of subjects administered some form of therapy and the group with no therapy reported. The hazard rate is very low when the subjects are administered any type of therapy

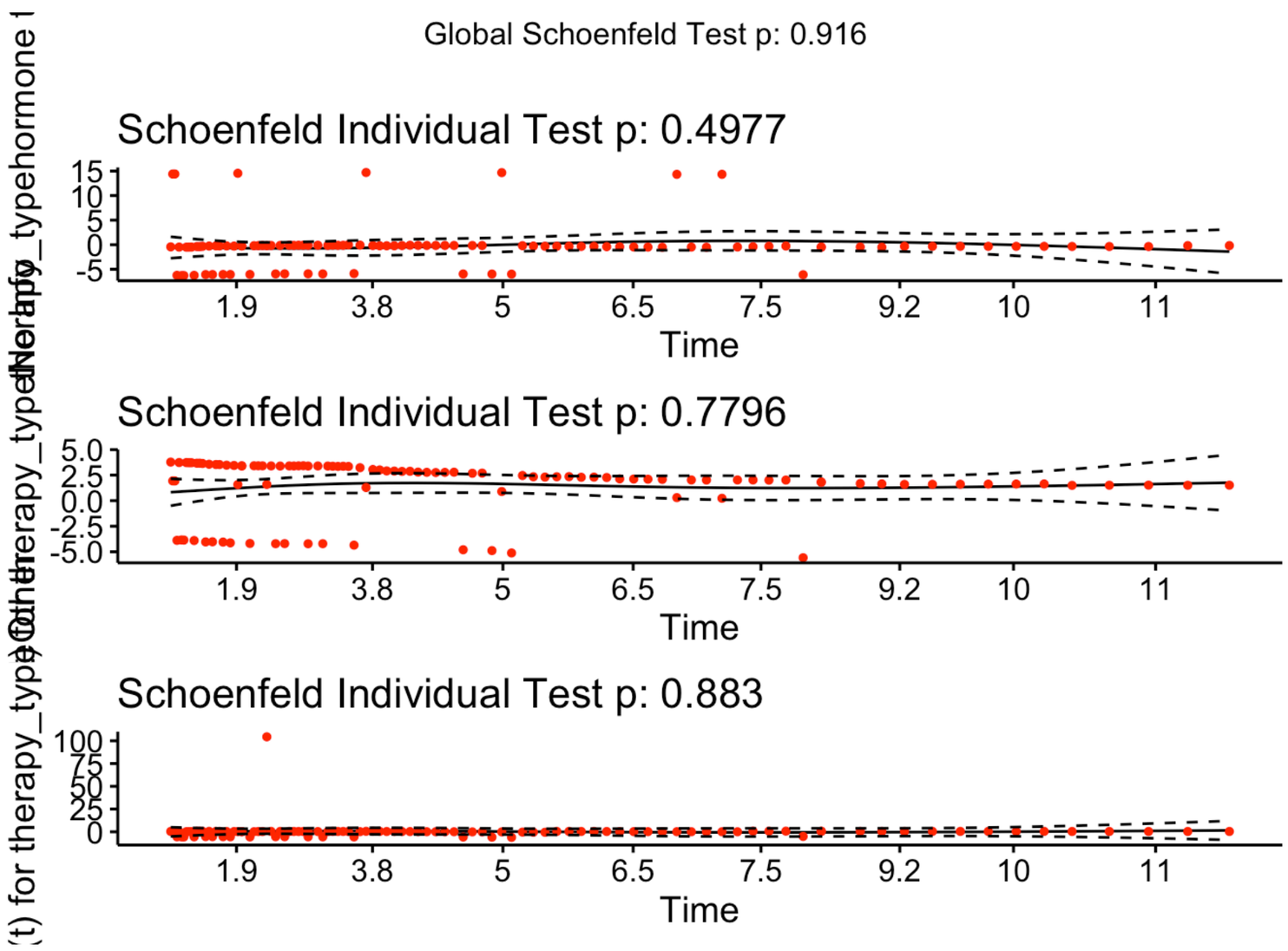
but the hazard rate is very high for subjects when no therapy is reported. The median survival rate is not even reached for the therapy related groups which indicates that more than half of the subjects are alive when the probability of survival is 50%. But, the median survival time is less than 7 years with no therapy.

When the survival curves are compared, it can be seen that the log-rank statistic or the MH statistic with a chi square distribution with a 3 degrees of freedom has a very high value of 44 and a p-value < 0.01. Using a confidence limit of 0.05, we have sufficient evidence to reject the null hypothesis and conclude that there is significant difference in the survival curves between the three groups. Further, by doing a pairwise comparison of the survival curves, the survival curve for the “No info” group is found to be significantly differnet from the curves for the other two groups. Basically, this indicates that the survival time is significantly different between subjects reciveing therapy and no therapy.

7.5.2 Cox Proportional Hazard (PH) Model for Therapy

Check for Proportional Hazards assumption

```
##                                 rho  chisq     p
## therapy_typehormone therapy  0.0668 0.4599 0.498
## therapy_typeNo Info          0.0278 0.0783 0.780
## therapy_typeOther           -0.0144 0.0216 0.883
## GLOBAL                          NA 0.5133 0.916
```

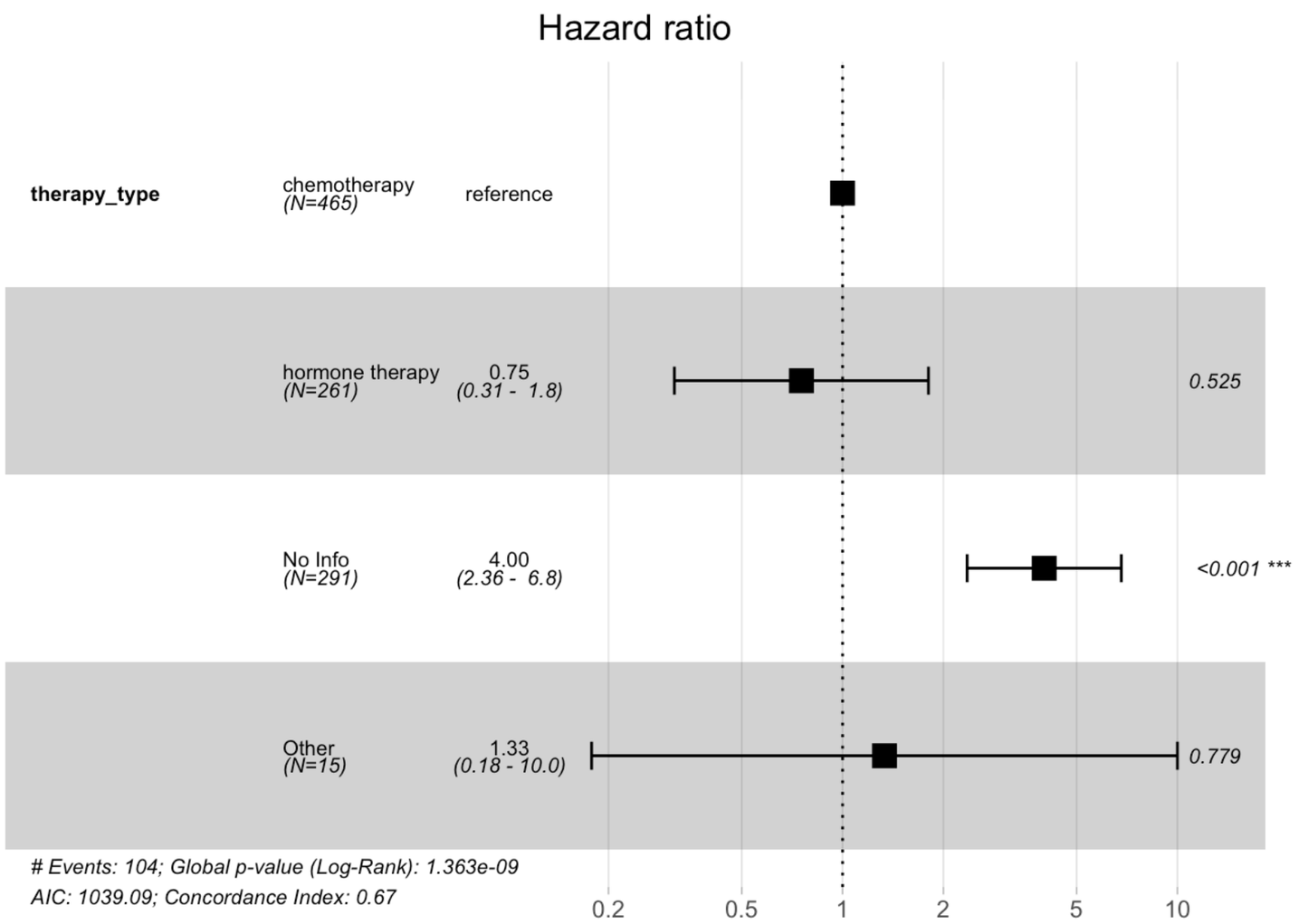


When tested for proportional hazards, the Schoenfeld residuals for all three treatment types are non-random around the mean line indicating that the hazards are proportional. This is also evident from the p-values which are all non-significant. This implies that the assumption of proportional hazards is satisfied.

Model Interpretation

```
## Call:
## coxph(formula = surv_obj_therapy ~ therapy_type, data = brca_clin_therapy)
##
##   n= 1032, number of events= 104
##
##                                coef exp(coef) se(coef)      z Pr(>|z|)
## therapy_typehormone therapy -0.2831    0.7535   0.4456 -0.635    0.525
## therapy_typeNo Info          1.3867    4.0015   0.2704  5.127 2.94e-07 ***
## therapy_typeOther            0.2881    1.3339   1.0276  0.280    0.779
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                             exp(coef) exp(-coef) lower .95 upper .95
## therapy_typehormone therapy    0.7535     1.3272    0.3146     1.805
## therapy_typeNo Info            4.0015     0.2499    2.3552     6.799
## therapy_typeOther              1.3339     0.7497    0.1780     9.997
##
## Concordance= 0.672  (se = 0.033 )
## Likelihood ratio test= 44.21  on 3 df,   p=1e-09
## Wald test            = 38.24  on 3 df,   p=3e-08
## Score (logrank) test = 44.01  on 3 df,   p=2e-09
```

```
## Warning: Removed 1 rows containing missing values (geom_errorbar).
```



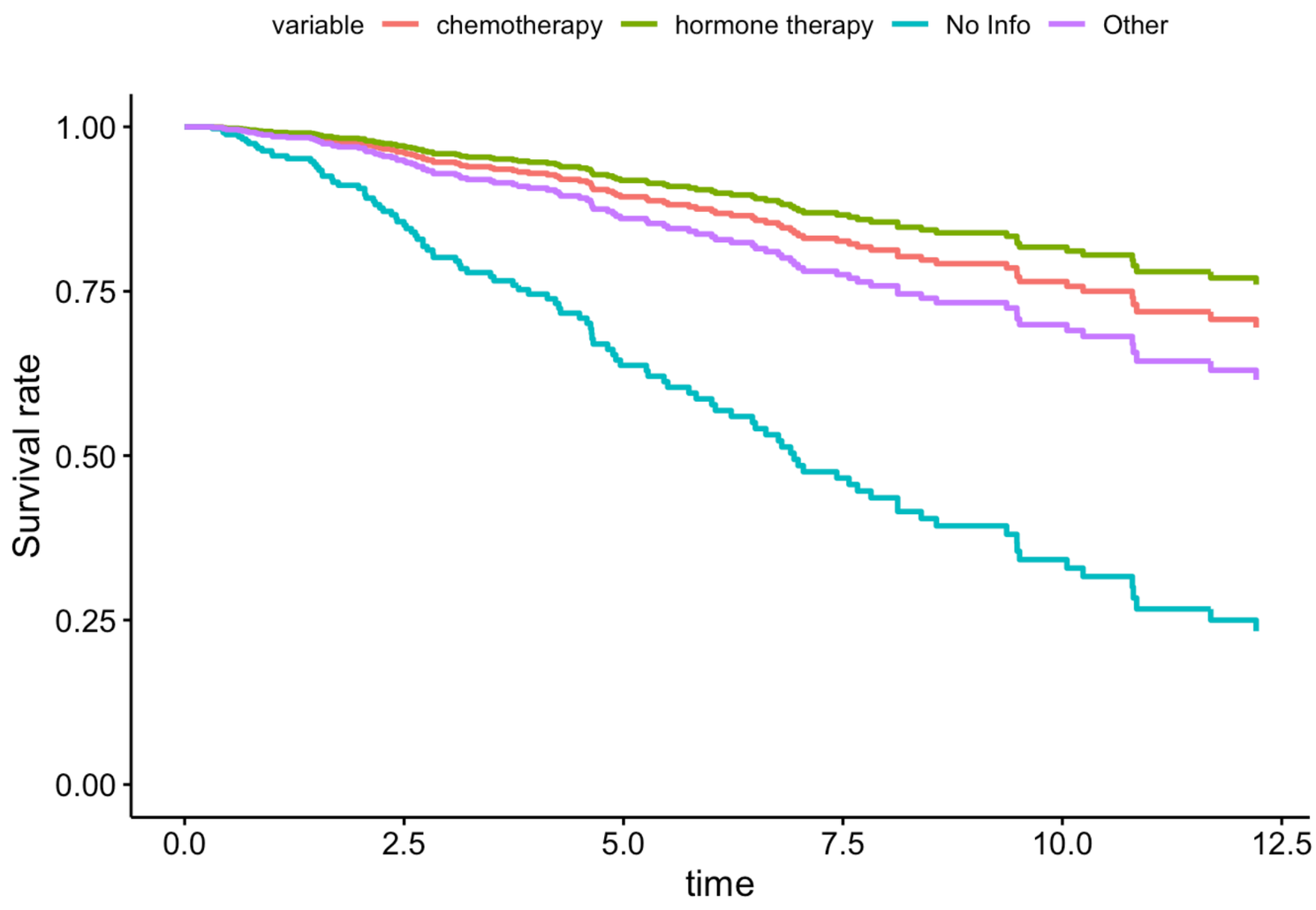
In this cox model where the treatment type is considered as a covariate, the “chemotherapy” treatment is considered as the baseline hazard. From the p-values for the Wald tests of the different coefficients, it is evident that the coefficient for only the “No-Info” group is significant. The Cox model equation can then be written as:

$$H(t|therapy) = \begin{cases} H_0(t) & therapy\_type = Chemotherapy, Hormone\ Therapy, Other \\ H_0(t)exp(1.3867) & therapy\_type = No\_Info \end{cases}$$

This equation can be used to compute the cumulative hazard rate for any therapy type.

Given a coefficient of 1.3867 and exp(coef)=4.0, the cumulative hazard rate for the “No Info” option is 4 times the hazard for the “chemotherapy” treatment. For the other treatments the cumulative hazard rate is the same as the one for chemotherapy.

Plots of Survival Curves by Therapy Type



7.5.3 Parametric modeling - Therapy

```
## Warning in survreg.fit(X, Y, weights, offset, init = init, controlvals =
## control, : Ran out of iterations and did not converge
```

```
## $vars
##                                  Estimate           SE
## lambda                         0.0002171904 5.133533e-05
## gamma                          3.5665853156 0.000000e+00
## therapy_typehormone therapy -0.2348434059 4.489286e-01
## therapy_typeNo Info           1.8014541337 2.406770e-01
## therapy_typeOther             0.7014825621 1.034854e+00
##
## $HR
##                                    HR        LB        UB
## therapy_typehormone therapy 0.7906947 0.3280047  1.906064
## therapy_typeNo Info         6.0584509 3.7800449  9.710156
## therapy_typeOther           2.0167404 0.2653263 15.329209
##
## $ETR
##                                   ETR        LB        UB
## therapy_typehormone therapy 1.0680616 0.8345550 1.3669029
## therapy_typeNo Info         0.6034500 0.5286902 0.6887812
## therapy_typeOther           0.8214519 0.4651614 1.4506432
```

The Weibull model is not a good fit as the p-value of the loglikelihood values of the model is high.

```
##
## Call:
## survreg(formula = surv_obj_therapy ~ therapy_type, data = brca_clin_therapy,
##     dist = "exponential")
##                               Value Std. Error     z       p
## (Intercept)                   4.006      0.236 17.00 < 2e-16
## therapy_typehormone therapy   0.300      0.445  0.67    0.50
## therapy_typeNo Info          -1.584      0.261 -6.06 1.4e-09
## therapy_typeOther            -0.345      1.027 -0.34    0.74
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -398.8   Loglik(intercept only)= -431.6
##  Chisq= 65.45 on 3 degrees of freedom, p= 4e-14
## Number of Newton-Raphson Iterations: 7
## n= 1032
```

```
## [1] "Constant Hazard Rate for Exponential Distribution with therapy as predictor: 0.02 events/year"
```

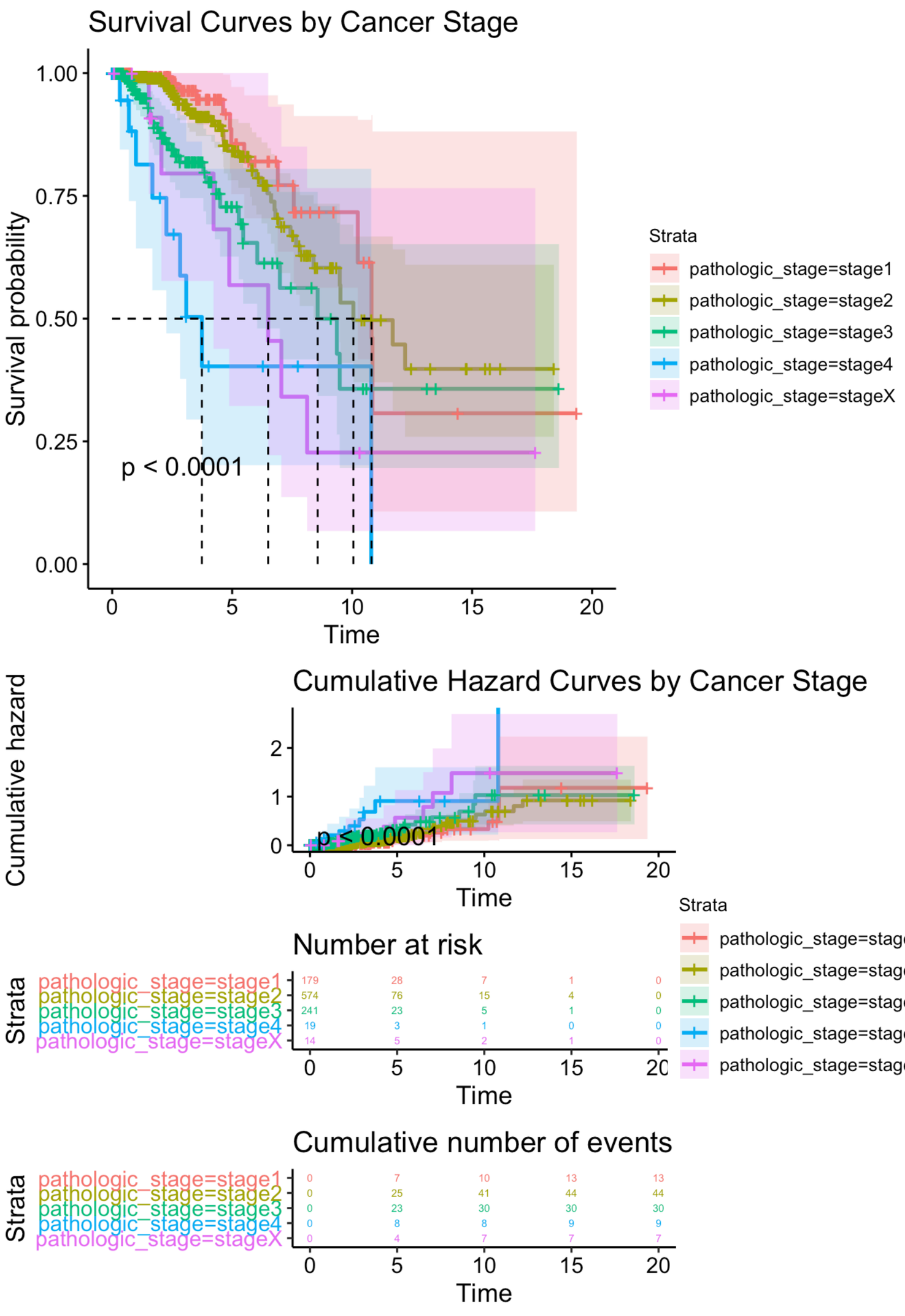
The exponential model is a good fit. Note that survreg() fits accelerated failure models, not proportional hazards models. The coefficients are logarithms of ratios of survival times, so a positive coefficient means longer survival.

From the p-values of the coefficients, we observe that only the coefficient of the No_Info group is significant. With a coefficient of 1.584, the survival time of the No_info group is increases by a factor of 0.2 (exp(-1.584)) or decreases by a factor of 0.8 as compared to the survival time for the chemotherapy group. Given that the hazard ratio for No_info vs chemotherapy group is found to be 4.87 (exp(-(-1.584))). Bascially, the risk of death for No_info group is almost 5 times higher than the risk of death for subjects given chemotherapy.

7.6 Cancer Stage (Cancer Stage)

7.6.1 Kaplan Meier Survival Curves for Cancer Stage

```
## Call: survfit(formula = surv_obj_stage ~ pathologic_stage, data = brca_clin_stage)
##
##                              n events median 0.95LCL 0.95UCL
## pathologic_stage=stage1 179     13  10.81   10.24      NA
## pathologic_stage=stage2 574     44  10.05    8.39      NA
## pathologic_stage=stage3 241     30   8.56    6.05      NA
## pathologic_stage=stage4  19      9   3.74    2.26      NA
## pathologic_stage=stageX  14      7   6.50    4.22      NA
```




```
## Call:
## survdiff(formula = surv_obj_stage ~ factor(pathologic_stage),
##     data = brca_clin_stage)
##
##                                  N Observed Expected (O-E)^2/E (O-E)^2/V
## factor(pathologic_stage)=stage1 179       13    21.61      3.43      4.35
## factor(pathologic_stage)=stage2 574       44    55.97      2.56      5.62
## factor(pathologic_stage)=stage3 241       30    19.41      5.78      7.14
## factor(pathologic_stage)=stage4  19        9     2.42     17.83     18.30
## factor(pathologic_stage)=stageX  14        7     3.58      3.27      3.41
##
##  Chisq= 32.9  on 4 degrees of freedom, p= 1e-06
```

```
##        stage1 stage2 stage3 stage4
## stage2
## stage3 **     **
## stage4 ****   ****   *
## stageX *      +
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

From the survival curves and the cumilative hazard curves, it can be observed that the hazard rate steadily increases with increase in the identified cancer stage number. Correspondingly, the median survival time steadily reduces from 10.8 years for stage 1 to 3.7 years for stage 4.

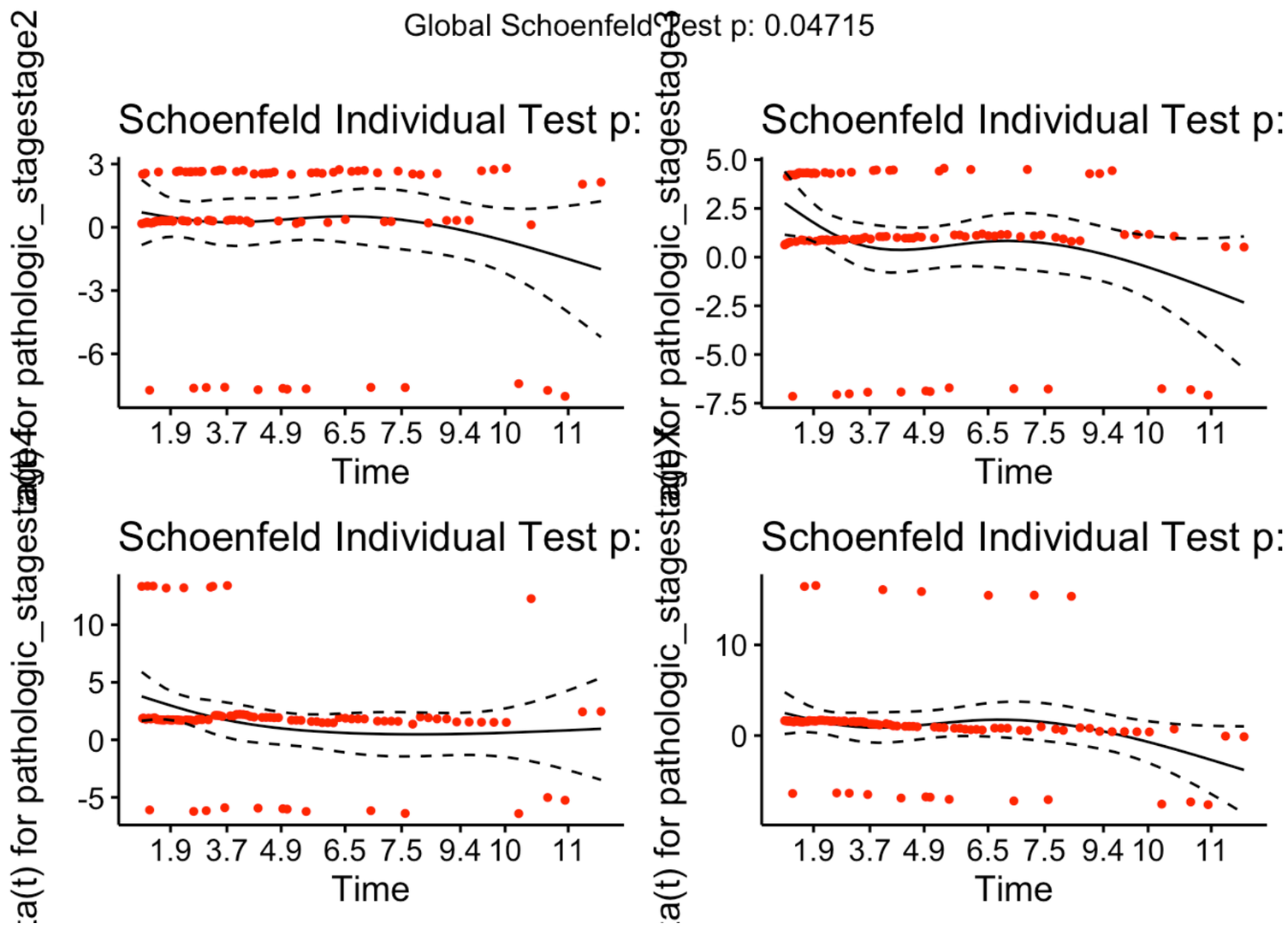
On comparison of the survival curves, it is found that the log rank statistic has a chi-square distribution with 4 degress of freedom and has a value of 32.9 and a very small p-value that is less than 0.01. Using a significance of 0.05 and with a p-value < 0.01, we have sufficient statistical evidence to reject the null hypothesis and conclude that the survival curves for the different stages are different from each other.

Further when the survival curves are compared with each other in pairs, it is found that only survival curves for stage1 and stage2 are not significantly different. All the other curves are significantly different from the survival curve for stage 1.

7.6.2 Cox Proportional Hazard (PH) Model for Pathologic Stage (Cancer Stage)

Check for Proportional Hazards Assumption

```
##                             rho chisq      p
## pathologic_stagestage2 -0.113  1.30 0.2537
## pathologic_stagestage3 -0.249  6.25 0.0124
## pathologic_stagestage4 -0.215  4.67 0.0307
## pathologic_stagestageX -0.166  2.87 0.0901
## GLOBAL                    NA  9.63 0.0472
```



When tested for proportional hazards, it seems that the hazard ratios of stage 3 and stage 4 with stage 1 are not proportional. We will attempt by splitting the survival data into multiple groups by time cutoff.

Split Survival Data into Time Groups

```
## Warning in cor(xx, r2): the standard deviation is zero
```

The split into multiple timegroups introduces other issues, like zero variance, into the Cox model and so we consider the Cox model to be not a good fit while using the cancer stage as a predictor.

7.6.3 Parametric modeling - Cancer Stage

```
##
## Call:
## survreg(formula = surv_obj_stage ~ pathologic_stage, data = brca_clin_stage,
##     dist = "weibull")
##                            Value Std. Error     z       p
## (Intercept)               2.9141     0.1886 15.45 < 2e-16
## pathologic_stagestage2 -0.1723     0.1992 -0.86  0.3871
## pathologic_stagestage3 -0.5875     0.2118 -2.77  0.0055
## pathologic_stagestage4 -1.1887     0.2826 -4.21 2.6e-05
## pathologic_stagestageX -0.6914     0.3047 -2.27  0.0233
## Log(scale)             -0.4619     0.0647 -7.14 9.6e-13
##
## Scale= 0.63
##
## Weibull distribution
## Loglik(model)= -395.2   Loglik(intercept only)= -407.6
##  Chisq= 24.71 on 4 degrees of freedom, p= 5.8e-05
## Number of Newton-Raphson Iterations: 16
## n= 1027
```

```
##
## Call:
## survreg(formula = surv_obj_stage ~ pathologic_stage, data = brca_clin_stage,
##     dist = "exponential")
##                          Value Std. Error     z       p
## (Intercept)              3.651      0.277 13.16 < 2e-16
## pathologic_stagestage2 -0.257      0.316 -0.81  0.4155
## pathologic_stagestage3 -0.868      0.332 -2.61  0.0090
## pathologic_stagestage4 -1.858      0.434 -4.28 1.8e-05
## pathologic_stagestageX -1.387      0.469 -2.96  0.0031
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -415.3   Loglik(intercept only)= -428
##  Chisq= 25.54 on 4 degrees of freedom, p= 3.9e-05
## Number of Newton-Raphson Iterations: 7
## n= 1027
```

```
## [1] "Constant Hazard Rate for Exponential Distribution with stage as predictor: 0.03  events/year"
```

The survival data is fit with parametric models using both the exponential and weibull distributions and compared. The fitness of the each of these models is verified by checking the loglikelihood values of the models. For both the models, p-values for the chi-square distribution with one degree of freedom are very small (<0.05) indicating that models are good fits.

```
## [1] "The better fit parametric model: weibull"
```

```
## $vars
##                             Estimate          SE
## lambda                   0.009804241 0.003279368
## gamma                    1.587102596 0.102729953
## pathologic_stagestage2 0.273380115 0.315675999
## pathologic_stagestage3 0.932479130 0.332135455
## pathologic_stagestage4 1.886664065 0.433696772
## pathologic_stagestageX 1.097250950 0.471224215
##
## $HR
##                                HR        LB        UB
## pathologic_stagestage2 1.314400 0.7079842  2.440233
## pathologic_stagestage3 2.540800 1.3251234  4.871747
## pathologic_stagestage4 6.597324 2.8197099 15.435871
## pathologic_stagestageX 2.995919 1.1896611  7.544610
##
## $ETR
##                                ETR        LB        UB
## pathologic_stagestage2 0.8417678 0.5697071 1.2437496
## pathologic_stagestage3 0.5556951 0.3668730 0.8417001
## pathologic_stagestage4 0.3046026 0.1750663 0.5299862
## pathologic_stagestageX 0.5008970 0.2756659 0.9101519
```

```
## [[1]]
##                             Estimate          SE
## lambda                   0.009804241 0.003279368
## gamma                    1.587102596 0.102729953
## pathologic_stagestage2 0.273380115 0.315675999
## pathologic_stagestage3 0.932479130 0.332135455
## pathologic_stagestage4 1.886664065 0.433696772
## pathologic_stagestageX 1.097250950 0.471224215
```

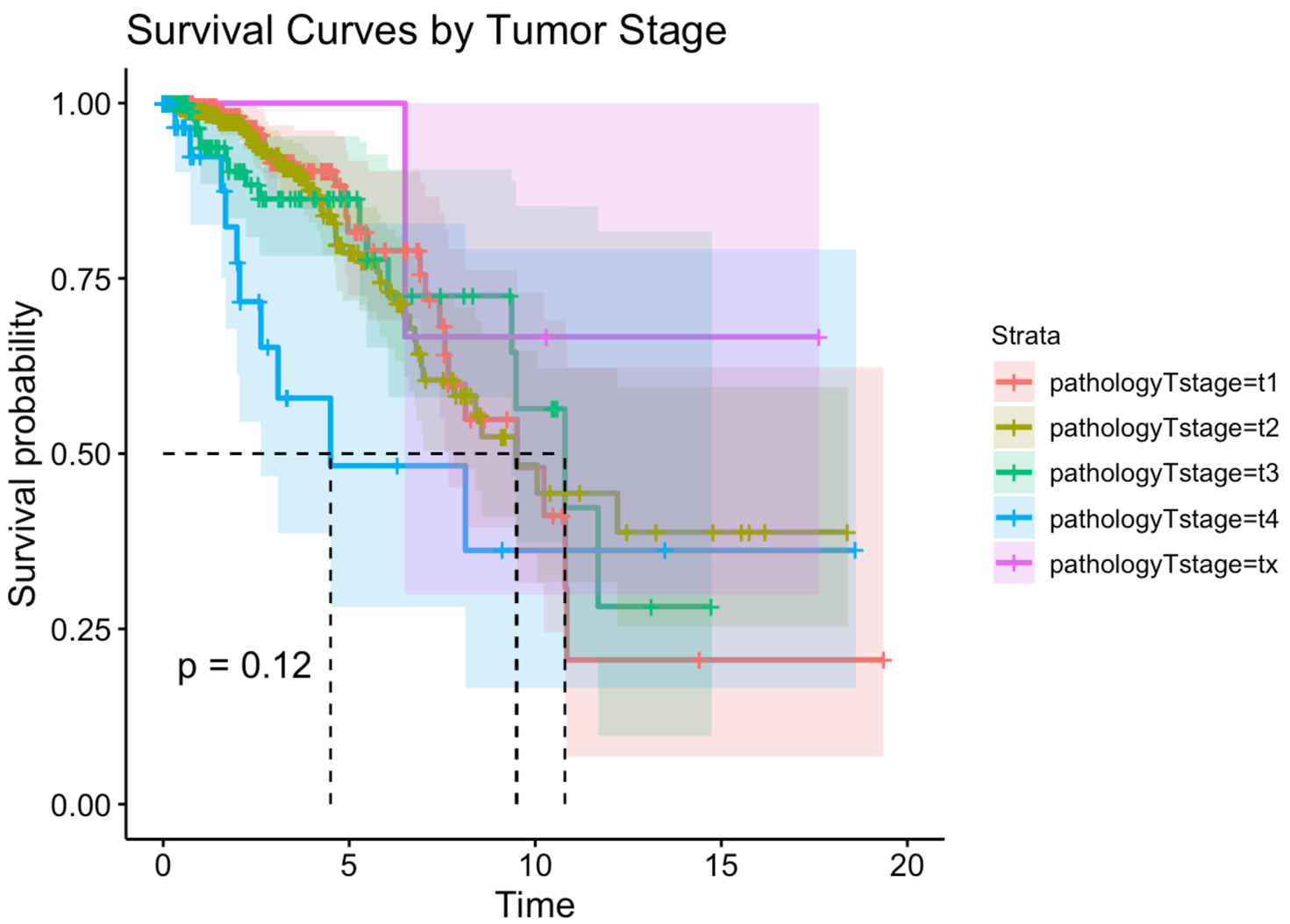
Anova is used to compare the two fitted models and it is found that the two model fits are significantly different and the model using assuming weibull distribution is the better one with the lower loglikelihood value.

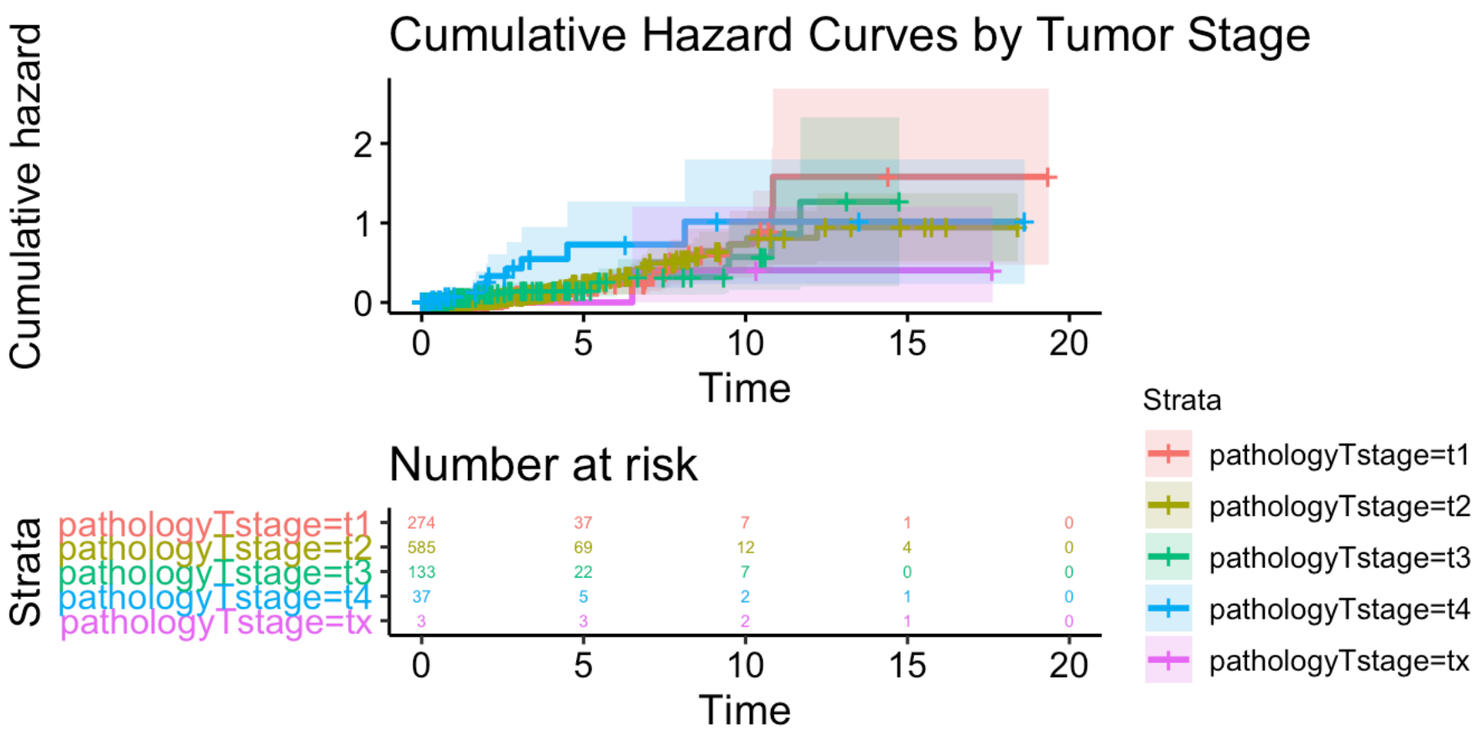
From the converted results of the weibull model, the risk of death increases with increase in the cancer stage number. When compared with the risk of death for stage 1, the risk is higher by a factor of 1.3 for stage 2, by a factor of 2.5 for stage 3 and by a factor of 6.6 for stage 4. Equivalently, the survival time for the stage 2 group reduces by 16% for stage 2, by 45% for stage 3 and by 70% for stage 4 when compared with the survival time length for stage 1.

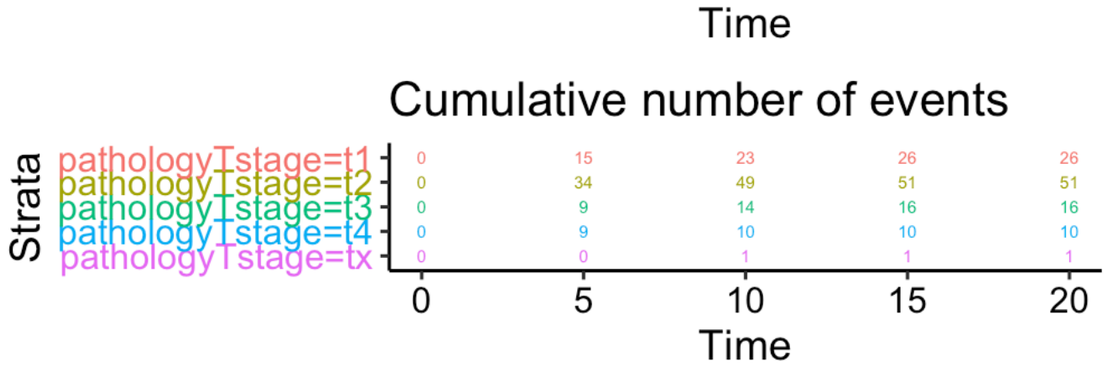
7.7 Tumor Stage (Pathology T)

7.7.1 Kaplan Meier Survival Curves for Tumor Stage

```
## Call: survfit(formula = surv_obj_Tstage ~ pathologyTstage, data = brca_clin_Tstage)
##
##                       n events median 0.95LCL 0.95UCL
## pathologyTstage=t1  274     26   9.51    7.67      NA
## pathologyTstage=t2  585     51   9.48    7.82      NA
## pathologyTstage=t3  133     16  10.80    9.36      NA
## pathologyTstage=t4   37     10   4.50    2.63      NA
## pathologyTstage=tx    3      1     NA    6.50      NA
```








```
## Call:
## survdiff(formula = surv_obj_Tstage ~ pathologyTstage, data = brca_clin_Tstage)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## pathologyTstage=t1  274       26    29.23  0.357443  0.500447
## pathologyTstage=t2  585       51    52.13  0.024402  0.049252
## pathologyTstage=t3  133       16    15.89  0.000796  0.000948
## pathologyTstage=t4   37       10     4.61  6.284557  6.643069
## pathologyTstage=tx    3        1     2.14  0.605271  0.627620
##
##  Chisq= 7.3  on 4 degrees of freedom, p= 0.1
```

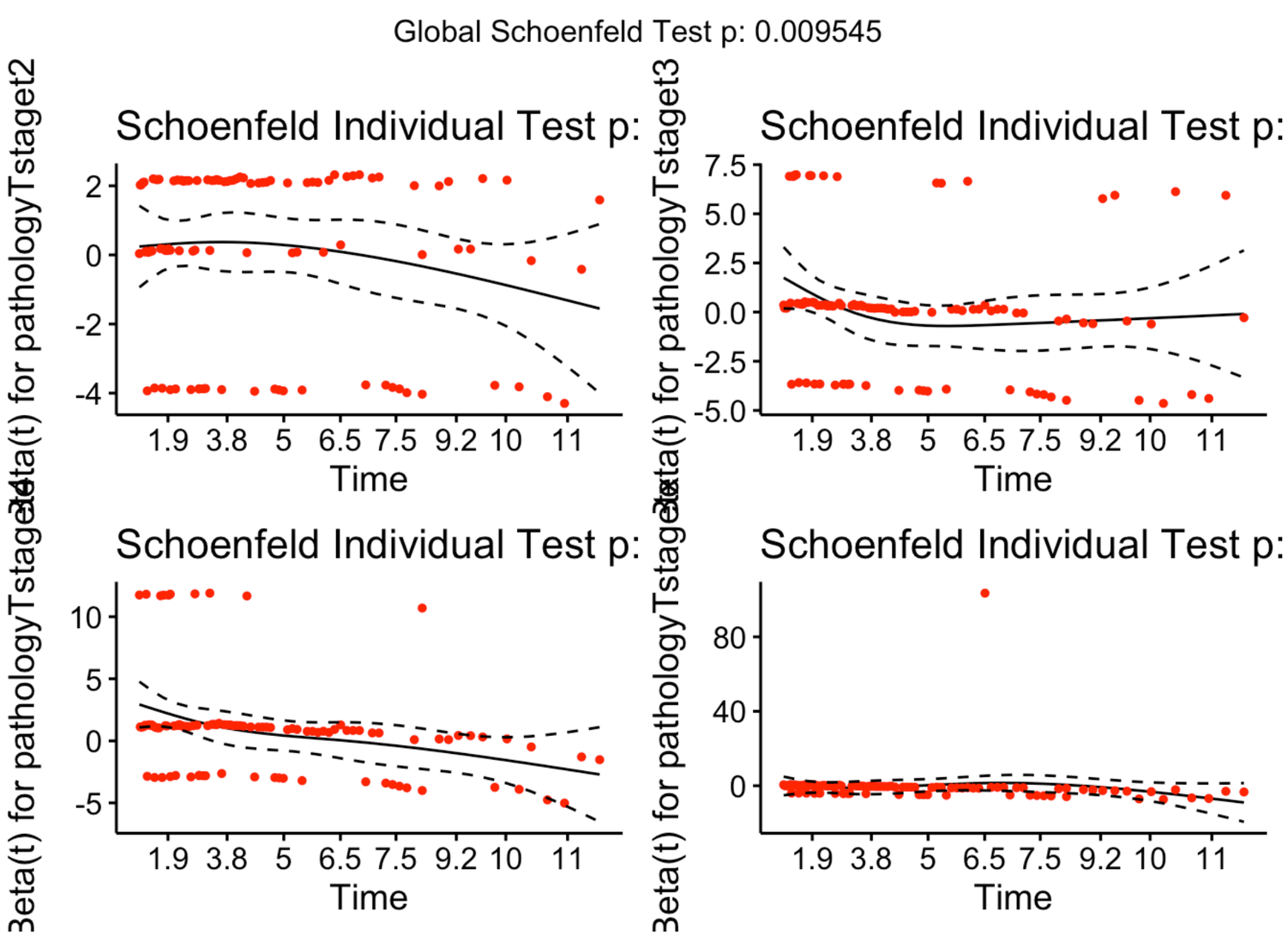
```
##    t1 t2 t3 t4
## t2
## t3
## t4 +  +
## tx
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

From the survival time and cumulative hazard plots, it can be observed that the probability of survival decreases gradually with time and increase in tumor classification type. The median survival time is an average of 10 years for the first three tumor stages and reduces to 4.5 years for the t4 stage. When the survival curves for the different tumor stages are compares, the difference in the curves is not found to be significantly different.

7.7.2 Cox Proportional Hazard (PH) Model for Pathologic Stage (T Stage)

Check for Proportional Hazards Assumption

```
##                        rho  chisq        p
## pathologyTstaget2 -0.1575  2.586 0.107807
## pathologyTstaget3 -0.1607  2.733 0.098307
## pathologyTstaget4 -0.3420 13.088 0.000297
## pathologyTstagetx -0.0614  0.395 0.529756
## GLOBAL                NA 13.384 0.009545
```



On checking the Cox model for the proportional hazards assumption, it is found that the hazard for stage 4 is not proportional. So, the Cox model is not a good fit as is. We split the survival data into multiple time groups.

Split the Survival Data into time groups

```
## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Loglik converged before variable 1,2,3,4 ; beta may be infinite.
```

```
## Warning in cor(xx, r2): the standard deviation is zero
```

The split into multiple timegroups introduces other issues, like zero variance, into the Cox model and so we do not consider the Cox model to be a good fit while using the tumor stage as a predictor.

7.7.3 Parametric modeling - Tstage

```
##
## Call:
## survreg(formula = surv_obj_Tstage ~ pathologyTstage, data = brca_clin_Tstage,
##     dist = "weibull")
##                      Value Std. Error     z       p
## (Intercept)         2.6581     0.1370 19.40 < 2e-16
## pathologyTstaget2  -0.0544     0.1519 -0.36   0.720
## pathologyTstaget3  -0.0962     0.2008 -0.48   0.632
## pathologyTstaget4  -0.4855     0.2424 -2.00   0.045
## pathologyTstagetx   0.5203     0.6422  0.81   0.418
## Log(scale)         -0.4623     0.0646 -7.16 8.3e-13
##
## Scale= 0.63
##
## Weibull distribution
## Loglik(model)= -407.9   Loglik(intercept only)= -410.4
##  Chisq= 5.14 on 4 degrees of freedom, p= 0.27
## Number of Newton-Raphson Iterations: 15
## n= 1032
```

```
##
## Call:
## survreg(formula = surv_obj_Tstage ~ pathologyTstage, data = brca_clin_Tstage,
##     dist = "exponential")
##                      Value Std. Error     z      p
## (Intercept)         3.2910     0.1961 16.78 <2e-16
## pathologyTstaget2  -0.0855     0.2410 -0.35 0.7228
## pathologyTstaget3  -0.2432     0.3177 -0.77 0.4440
## pathologyTstaget4  -1.0100     0.3721 -2.71 0.0066
## pathologyTstagetx   0.2486     1.0190  0.24 0.8073
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -428.2   Loglik(intercept only)= -431.6
##  Chisq= 6.81 on 4 degrees of freedom, p= 0.15
## Number of Newton-Raphson Iterations: 7
## n= 1032
```

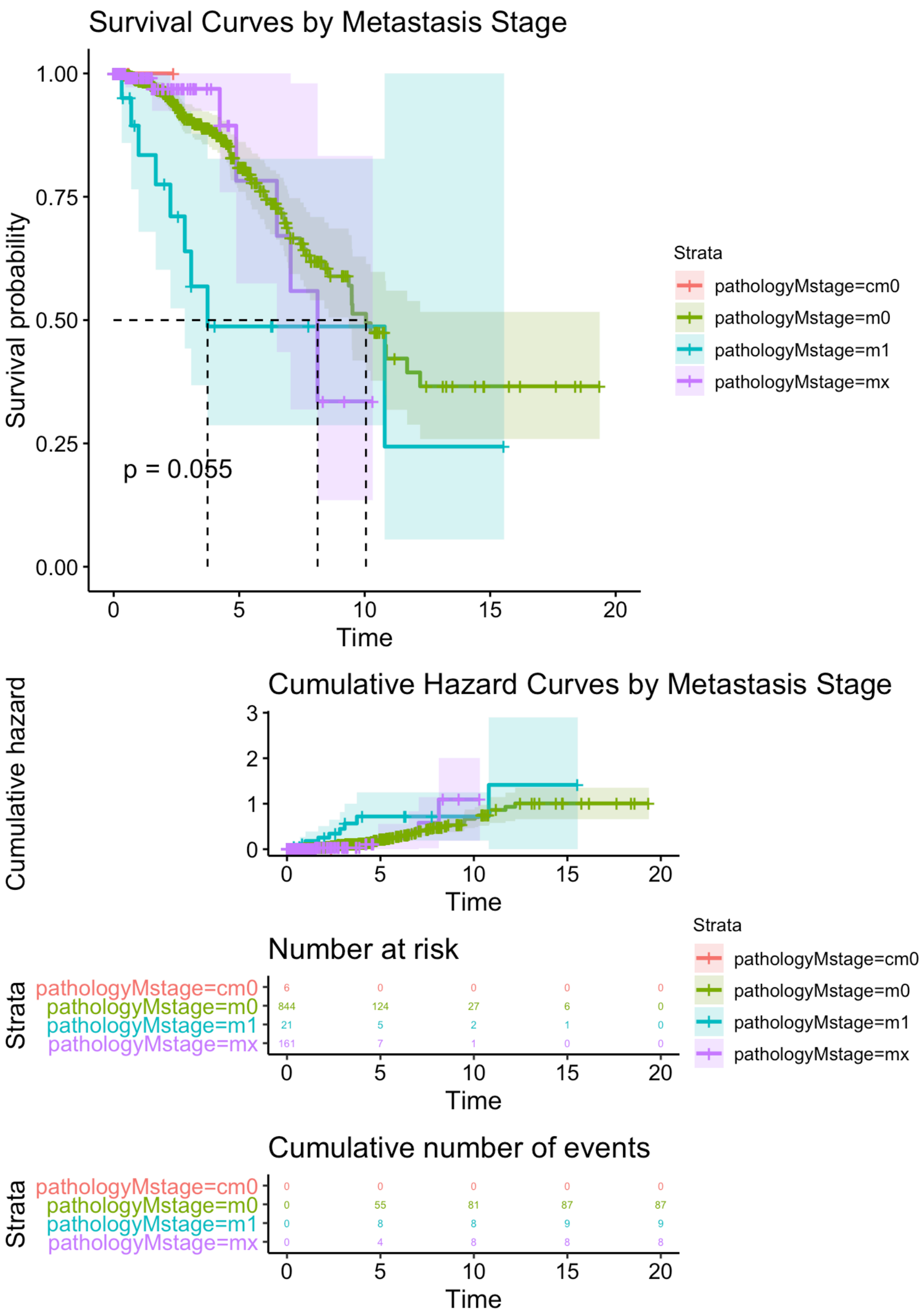
```
## [1] "Constant Hazard Rate for Exponential Distribution with Tstage as predictor: 0.04  events/year"
```

The survival data is fit with parametric models using both the exponential and weibull distributions. The fitness of the each of these models is first verified by checking the loglikelihood values of the models. For both the models, p-values for the chi-square distribution with one degree of freedom are high (> 0.05) indicating that both the parametric models are not good fits.

7.8 Metastasis Status (Pathology M)

7.8.1 Kaplan Meier Survival Curves for Metastasis Stage

```
## Call: survfit(formula = surv_obj_Mstage ~ pathologyMstage, data = brca_clin_Mstage)
##
##                         n events median 0.95LCL 0.95UCL
## pathologyMstage=cm0     6      0     NA      NA      NA
## pathologyMstage=m0    844     87  10.05    9.36      NA
## pathologyMstage=m1     21      9   3.74    2.83      NA
## pathologyMstage=mx    161      8   8.12    6.50      NA
```




```
## Call:
## survdiff(formula = surv_obj_Mstage ~ factor(pathologyMstage),
##     data = brca_clin_Mstage)
##
##                                N Observed Expected (O-E)^2/E (O-E)^2/V
## factor(pathologyMstage)=cm0    6        0   0.0805   0.08048    0.0807
## factor(pathologyMstage)=m0   844       87  92.3632   0.31142    2.7987
## factor(pathologyMstage)=m1    21        9   3.7855   7.18305    7.4905
## factor(pathologyMstage)=mx   161        8   7.7708   0.00676    0.0074
##
##  Chisq= 7.6  on 3 degrees of freedom, p= 0.05
```

```
##    cm0 m0 m1
## m0
## m1     *
## mx        +
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

From the cumulative hazard curves,it can be observed that for patients with no metastasis (M0 and CM0 stages), the hazard rate is very low for the initial five years and then linearly increases with time and becomes steady after 12 years. The median survival time in this case is around 10 years. When there is metastasis, the hazard rate is very high during the first five years and then becomes steady. The median survival time reduces to 3.7 years in this case.

But when the survival curves for the different stages are statistically compared, the log rank statistic for a chi-square distribution with 2 degress of freedom has a value of 7.6 and a p-value of 0.05. Using a significance of 0.05, we do not have very strong statistical evidence to reject the null hypothesis and conclude that the survival curves for the different stages may not be signifcantly different from each other. Still, when the survival curves are compared with each other using a family-wise p-

value, the survival curves for stagem0 and m1 are found to be different.

7.8.2 Cox Proportional Hazard (PH) Model for Pathologic Stage (M Stage)

```
## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Loglik converged before variable 1,2,3 ; coefficient may be infinite.
```

Fitting the Cox model with the metastasis stage as the predictor introduces inifinite values and so the Cox model is not considered as a good fit.

7.8.3 Parametric modeling - Mstage

```
##
## Call:
## survreg(formula = surv_obj_Mstage ~ pathologyMstage, data = brca_clin_Mstage,
##     dist = "weibull")
##                       Value Std. Error       z      p
## (Intercept)         56.5814     0.2345  241.25 <2e-16
## pathologyMstagem0 -53.9515     0.2340 -230.61 <2e-16
## pathologyMstagem1 -54.5237     0.3153 -172.92 <2e-16
## pathologyMstagemx -54.0256     0.0000    -Inf <2e-16
## Log(scale)         -0.4612     0.0647   -7.13  1e-12
##
## Scale= 0.631
##
## Weibull distribution
## Loglik(model)= -407.7   Loglik(intercept only)= -410.4
##  Chisq= 5.48 on 3 degrees of freedom, p= 0.14
## Number of Newton-Raphson Iterations: 14
## n= 1032
```

```
##
## Call:
## survreg(formula = surv_obj_Mstage ~ pathologyMstage, data = brca_clin_Mstage,
##     dist = "exponential")
##                     Value Std. Error     z p
## (Intercept)          17.5     2861.9  0.01 1
## pathologyMstagem0   -14.3     2861.9  0.00 1
## pathologyMstagem1   -15.3     2861.9 -0.01 1
## pathologyMstagemx   -14.1     2861.9  0.00 1
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -427.7   Loglik(intercept only)= -431.6
##  Chisq= 7.7 on 3 degrees of freedom, p= 0.053
## Number of Newton-Raphson Iterations: 17
## n= 1032
```

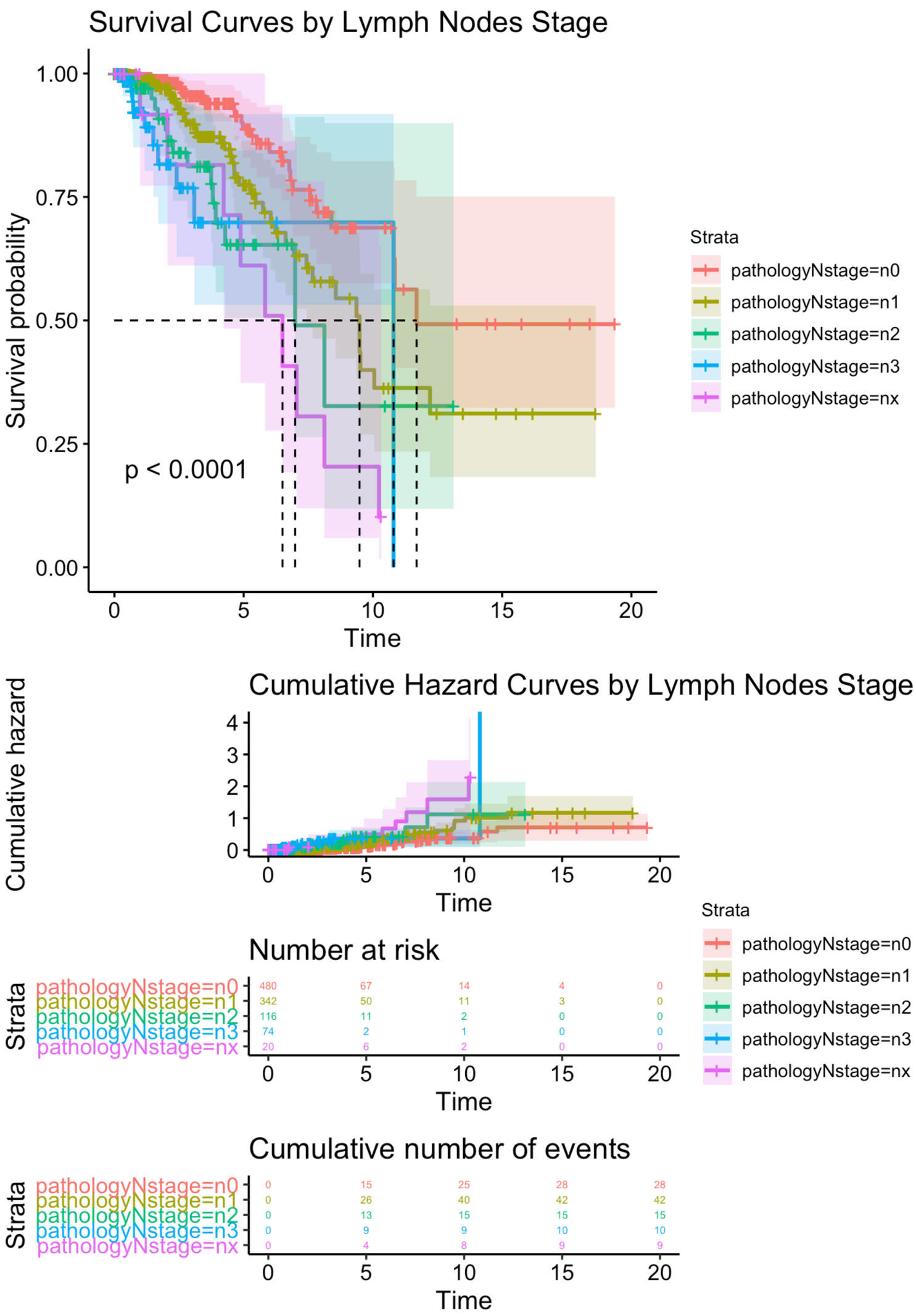
```
## [1] "Constant Hazard Rate for Exponential Distribution with Mstage as predictor: 0  events/year"
```

The survival data is fit with parametric models using both the exponential and weibull distributions and compared. The fitness of the each of these models is verified by checking the loglikelihood values of the models. For both the models, p-values for the chi-square distribution with one degree of freedom are high (> 0.05) indicating that models are not good fits.

7.9 Lymph Nodes Status (Pathology N)

7.9.1 Kaplan Meier Survival Curves for Lymph Nodes Stage

```
## Call: survfit(formula = surv_obj_Nstage ~ pathologyNstage, data = brca_clin_Nstage)
##
##                       n events median 0.95LCL 0.95UCL
## pathologyNstage=n0  480     28  11.69   10.81      NA
## pathologyNstage=n1  342     42   9.48    7.43      NA
## pathologyNstage=n2  116     15   6.99    6.99      NA
## pathologyNstage=n3   74     10  10.80      NA      NA
## pathologyNstage=nx   20      9   6.50    4.22      NA
```




```
## Call:
## survdiff(formula = surv_obj_Nstage ~ factor(pathologyNstage),
##     data = brca_clin_Nstage)
##
##                               N Observed Expected (O-E)^2/E (O-E)^2/V
## factor(pathologyNstage)=n0 480       28    49.88     9.596    18.509
## factor(pathologyNstage)=n1 342       42    38.42     0.333     0.529
## factor(pathologyNstage)=n2 116       15     8.59     4.792     5.249
## factor(pathologyNstage)=n3  74       10     3.50    12.062    12.626
## factor(pathologyNstage)=nx  20        9     3.61     8.045     8.406
##
##  Chisq= 35.1  on 4 degrees of freedom, p= 4e-07
```

```
##    n0   n1 n2 n3
## n1 *
## n2 ***
## n3 **** *
## nx ***  +
## attr(,"legend")
## [1] 0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 \t    ## NA: ''
```

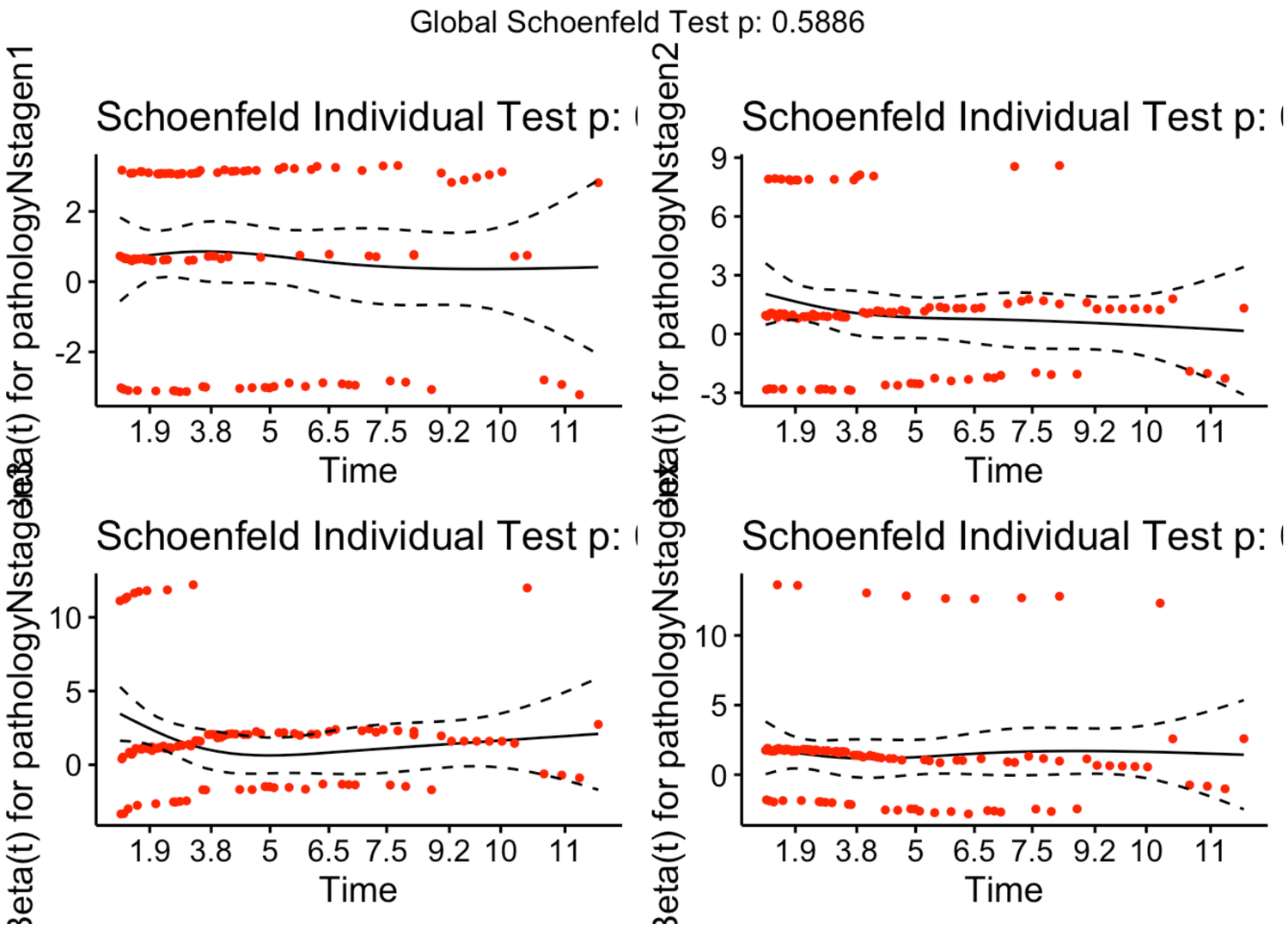
From the survival and cumulative hazard curves, it can be observed that the hazard rates increase with increase in the number of lymph nodes (represented by the stage numbers) that the tumor cells are spread to. For the first years, the hazard rates are very similar for the different stages but differ greatly after that. The median survival time decreases from 11.7 to 6.5 from stage N0 to stage NX.

When the survival curves for the different N stages are compared, the log-rank or MH_statistic for a chi-square distribution with 4 degrees of freedom is 35.1 with a very small p-value. So, we have sufficient statistical evidence to reject the null hypothesis that the survival curves are all the same and conclude that the survival curves are significantly different for the different N stages. When the survival curves are compared with each other in pairs, it is found that the survival curve for stage n0 is significantly different from the curves for all other stages.

7.9.2 Cox Proportional Hazard (PH) Model for Pathologic Stage (N Stage)

Check for Proportional Hazards Assumption

```
##                          rho    chisq     p
## pathologyNstagen1 -0.05455 3.03e-01 0.582
## pathologyNstagen2 -0.14261 2.06e+00 0.152
## pathologyNstagen3 -0.10898 1.12e+00 0.289
## pathologyNstagenx  0.00086 7.48e-05 0.993
## GLOBAL                NA 2.82e+00 0.589
```



When tested for proportional hazards, the Schoenfeld residuals for all four metastasis stages are non-random around the mean line indicating that the hazards are proportional. This is also evident from the p-values which are all non-significant. This implies that the proportional hazards condition is met.

Interpretation of the Model

```
## Call:
## coxph(formula = surv_obj_Nstage ~ pathologyNstage, data = brca_clin_Nstage)
##
##   n= 1032, number of events= 104
##
##                      coef exp(coef) se(coef)     z Pr(>|z|)
## pathologyNstagen1 0.6687    1.9517   0.2442 2.739 0.006166 **
## pathologyNstagen2 1.1473    3.1498   0.3212 3.572 0.000354 ***
## pathologyNstagen3 1.6542    5.2288   0.3737 4.427 9.56e-06 ***
## pathologyNstagenx 1.4943    4.4564   0.3860 3.871 0.000108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## pathologyNstagen1     1.952     0.5124     1.209     3.149
## pathologyNstagen2     3.150     0.3175     1.678     5.911
## pathologyNstagen3     5.229     0.1912     2.514    10.876
## pathologyNstagenx     4.456     0.2244     2.091     9.496
##
## Concordance= 0.675  (se = 0.034 )
## Likelihood ratio test= 29.61  on 4 df,   p=6e-06
## Wald test            = 31.05  on 4 df,   p=3e-06
## Score (logrank) test = 35.14  on 4 df,   p=4e-07
```

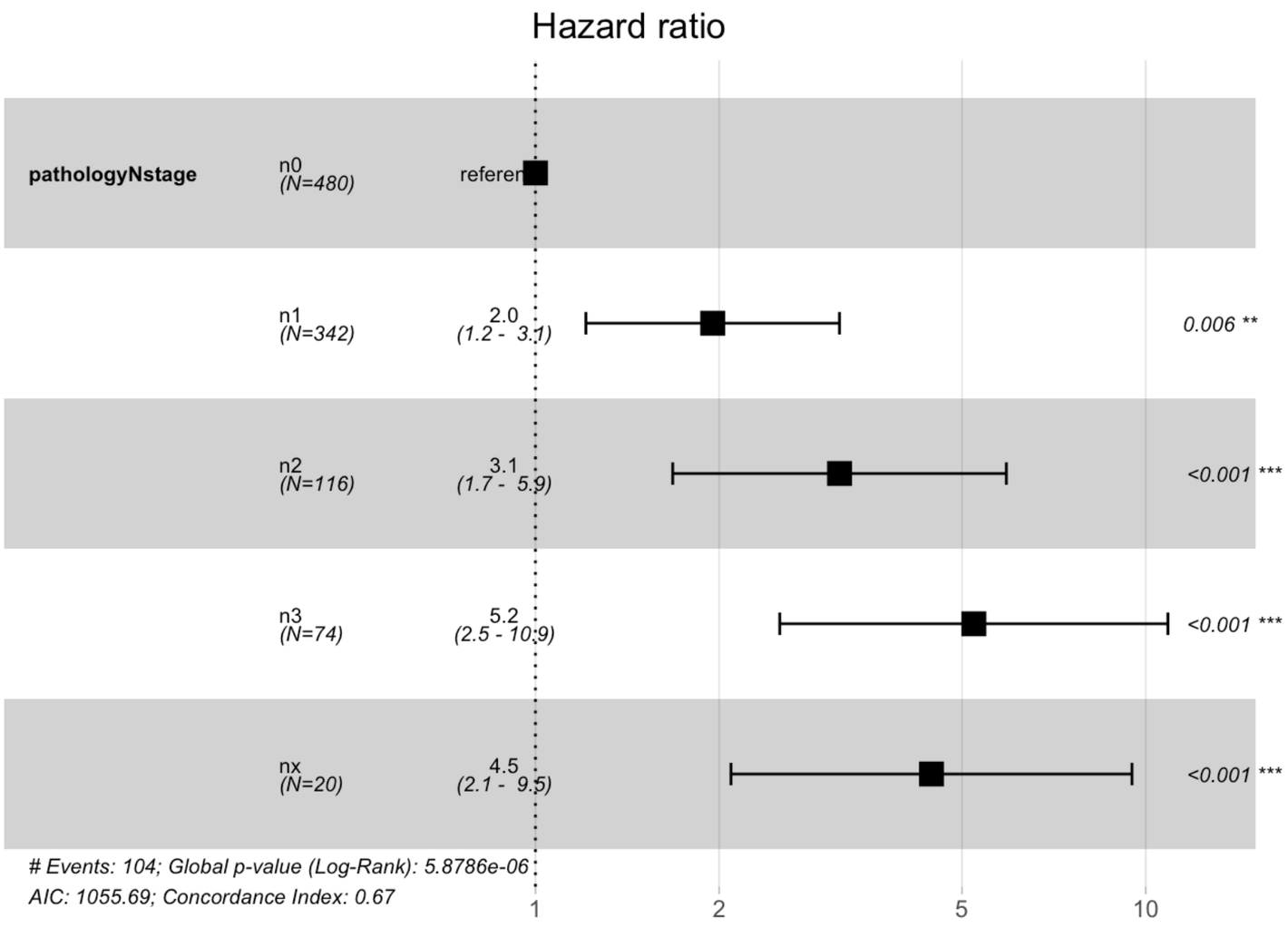
In this Cox model where the N stage is considered as a covariate, stage 0 (no metastasis to lymph nodes) is considered as the baseline hazard. From the p-values for the Wald tests of the different coefficients, it is evident that all the coefficients are significant. The Cox model equation can then be written as:

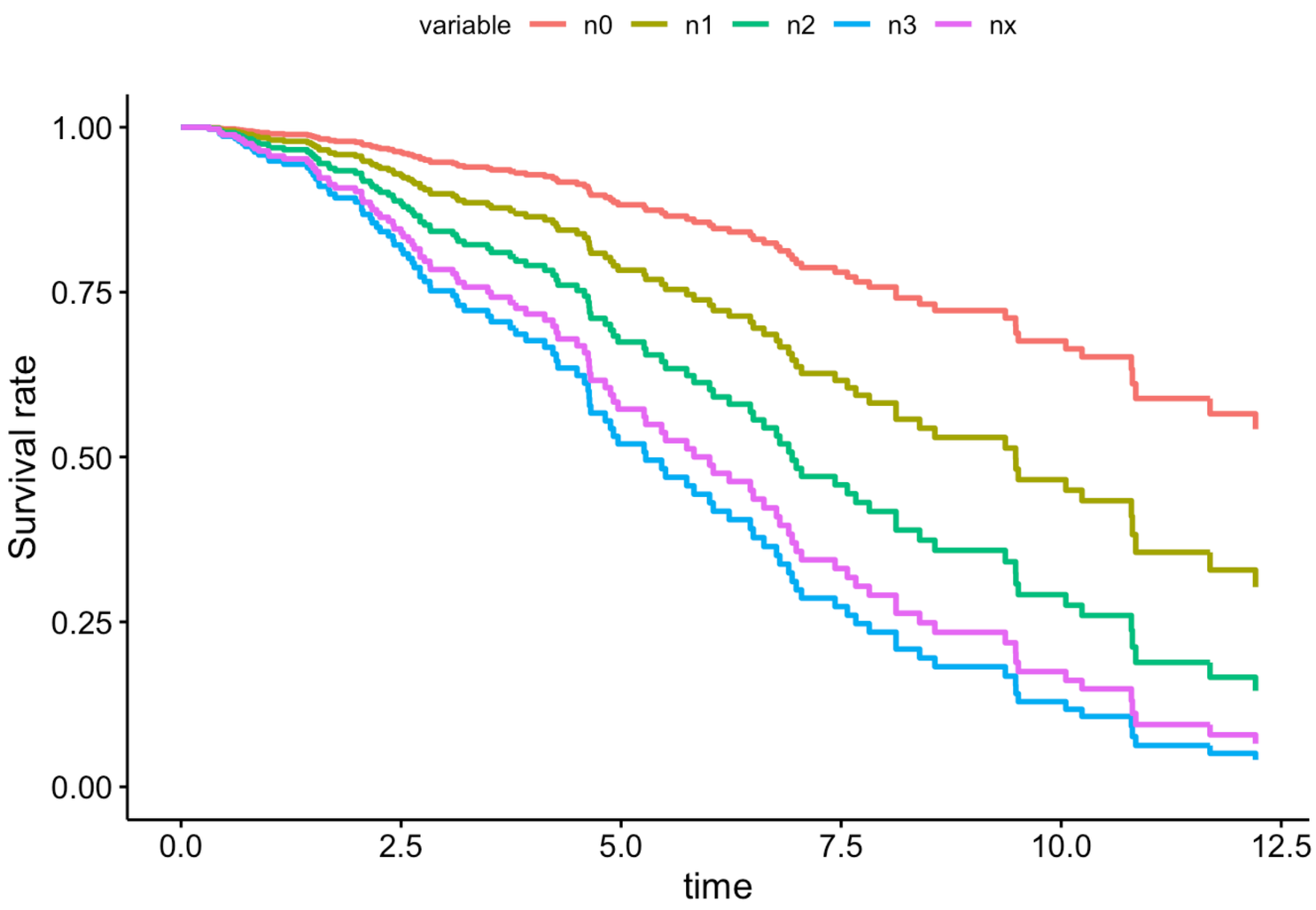
$$H(t|n0, n1, n2, n3, nX) = H_0(t)exp(0.6687 * n1 + 1.1473 * n2 + 1.6542 * n3 + 1.4943 * nX)$$

The cumulative hazard and hence the Survival function can be estimated from this equation for any new subject.

The relative risk of death can be estimated from the hazard ratios (exp(coef)) for the different stages. When compared to the group of subjects with no metastasis in the lymph nodes (stage n0), the risk of death increases by a factor of 1.95 when the metastasis increases to 1-3 lymph nodes (stage n1), increases by a factor of 3.15 when the metastasis increases to 4-9 lymph nodes (stage n2) and by a factor of 5.22 when metastasis increases to 10 or more lymph nodes (stage n3). So, the pathologyNstage is a strong predictor. This can also be seen from the plot of the hazard ratio and the plot of surival rates.

```
## Warning: Removed 1 rows containing missing values (geom_errorbar).
```





7.9.3 Parametric modeling - Nstage

```
##
## Call:
## survreg(formula = surv_obj_Nstage ~ pathologyNstage, data = brca_clin_Nstage,
##     dist = "weibull")
##                     Value Std. Error     z       p
## (Intercept)        2.9191     0.1315 22.20 < 2e-16
## pathologyNstagen1 -0.4148     0.1493 -2.78 0.00547
## pathologyNstagen2 -0.7205     0.1954 -3.69 0.00023
## pathologyNstagen3 -1.0281     0.2244 -4.58 4.6e-06
## pathologyNstagenx -0.9640     0.2408 -4.00 6.3e-05
## Log(scale)        -0.5043     0.0643 -7.84 4.6e-15
##
## Scale= 0.604
##
## Weibull distribution
## Loglik(model)= -394.4   Loglik(intercept only)= -410.4
##  Chisq= 32.07 on 4 degrees of freedom, p= 1.8e-06
## Number of Newton-Raphson Iterations: 15
## n= 1032
```

```
##
## Call:
## survreg(formula = surv_obj_Nstage ~ pathologyNstage, data = brca_clin_Nstage,
##     dist = "exponential")
##                     Value Std. Error     z       p
## (Intercept)         3.714      0.189 19.65 < 2e-16
## pathologyNstagen1 -0.671      0.244 -2.75 0.00597
## pathologyNstagen2 -1.036      0.320 -3.24 0.00120
## pathologyNstagen3 -1.313      0.368 -3.56 0.00036
## pathologyNstagenx -1.706      0.383 -4.45 8.5e-06
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -418.1   Loglik(intercept only)= -431.6
##  Chisq= 26.93 on 4 degrees of freedom, p= 2.1e-05
## Number of Newton-Raphson Iterations: 7
## n= 1032
```

```
## [1] "Constant Hazard Rate for Exponential Distribution with Nstage as predictor: 0.02  events/year"
```

The survival data is fit with parametric models using both the exponential and weibull distributions and compared. The fitness of the each of these models is verified by checking the loglikelihood values of the models. For both the models, p-values for the chi-square distribution with one degree of freedom are very small (<0.05) indicating that models are good fits.

```
## [1] "The better fit parametric model: weibull"
```

```
## $vars
##                      Estimate          SE
## lambda            0.007956353 0.002206135
## gamma             1.655887964 0.106539769
## pathologyNstagen1 0.686919158 0.244007708
## pathologyNstagen2 1.193129477 0.321238478
## pathologyNstagen3 1.702385303 0.373425326
## pathologyNstagenx 1.596287400 0.383256511
##
## $HR
##                          HR       LB         UB
## pathologyNstagen1 1.987583 1.232042  3.206453
## pathologyNstagen2 3.297384 1.756835  6.188825
## pathologyNstagen3 5.487020 2.639224 11.407665
## pathologyNstagenx 4.934678 2.328254 10.458932
##
## $ETR
##                          ETR        LB        UB
## pathologyNstagen1 0.6604497 0.4928721 0.8850040
## pathologyNstagen2 0.4864907 0.3316926 0.7135316
## pathologyNstagen3 0.3576931 0.2304142 0.5552798
## pathologyNstagenx 0.3813617 0.2378730 0.6114052
```

```
## [[1]]
##                       Estimate          SE
## lambda             0.007956353 0.002206135
## gamma              1.655887964 0.106539769
## pathologyNstagen1  0.686919158 0.244007708
## pathologyNstagen2  1.193129477 0.321238478
## pathologyNstagen3  1.702385303 0.373425326
## pathologyNstagenx  1.596287400 0.383256511
```

Anova is used to compare the two fitted models and it is found that the two model fits are significantly different with the weibull parametric being the better one with the lower loglikelihood value.

From the converted results of the weibull model, when compared to the group of subjects with no metastasis in the lymph nodes, the risk of death increases when the metastasis increases to 1-3 lymph nodes, increases by 300% when the metastasis increases to 4-9 lymph nodes and by 550% whe metastasis increases to 10 or more lymph nodes. Equivalently, when compared with no metastatis, the survival time decreases by 34% when for 1-3 lymph nodes, by 52% for 4-9 lymph nodes and by 65% for metastasis in 10 or more lymph nodes

7.10 Cox Proportional Hazard Model for All - no interactions

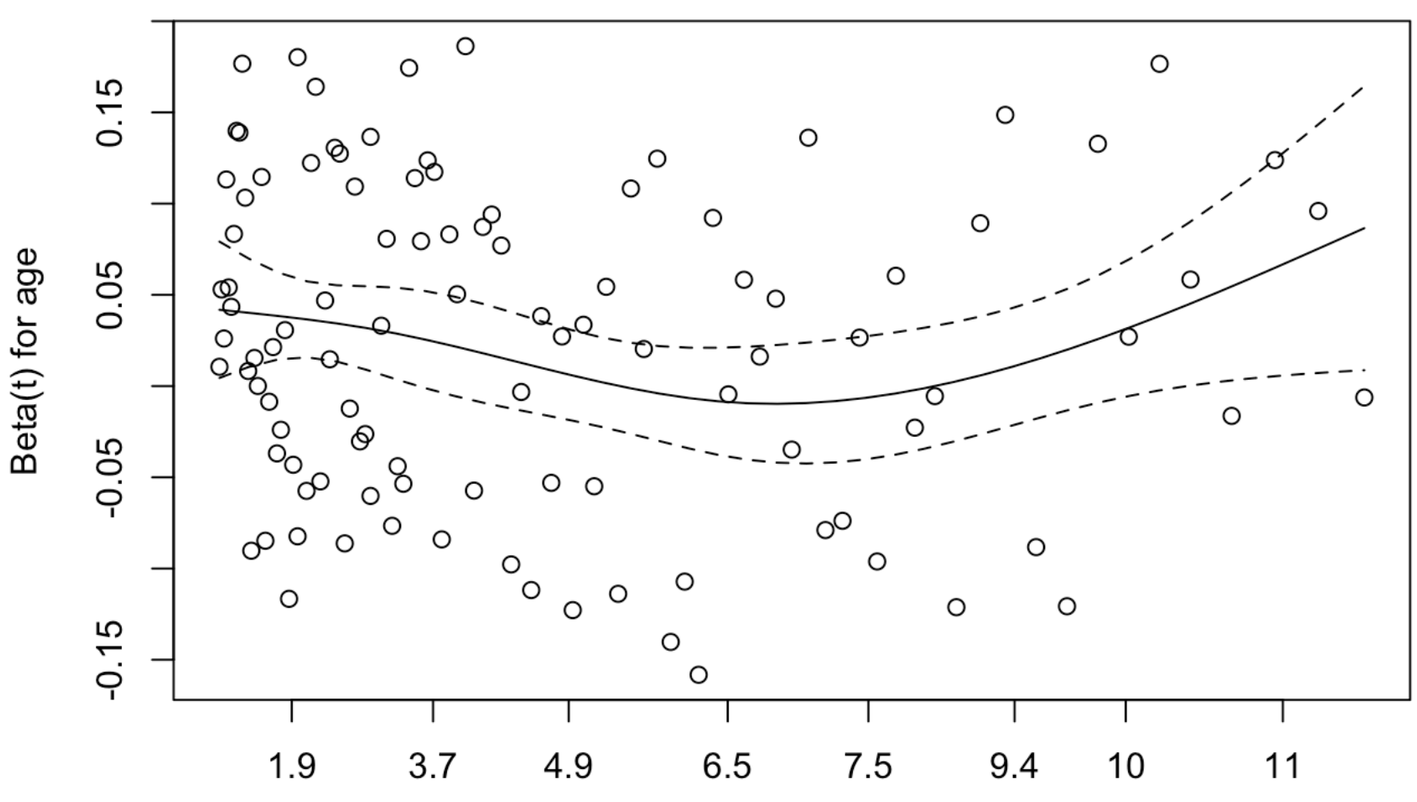
The final step is to consider all the predictors to determine the effect of each predictor on the hazard rate, while accounting for all other predictors. We had considered nine predictors - age at first diagnosis, race, ethnicity, therapy_type, cancer stage, tumor stage, metastasis stage, and lymph node stage. Of these race, ethnicity, and metastasis stage were not found to be significant. So, we drop them from our final list.

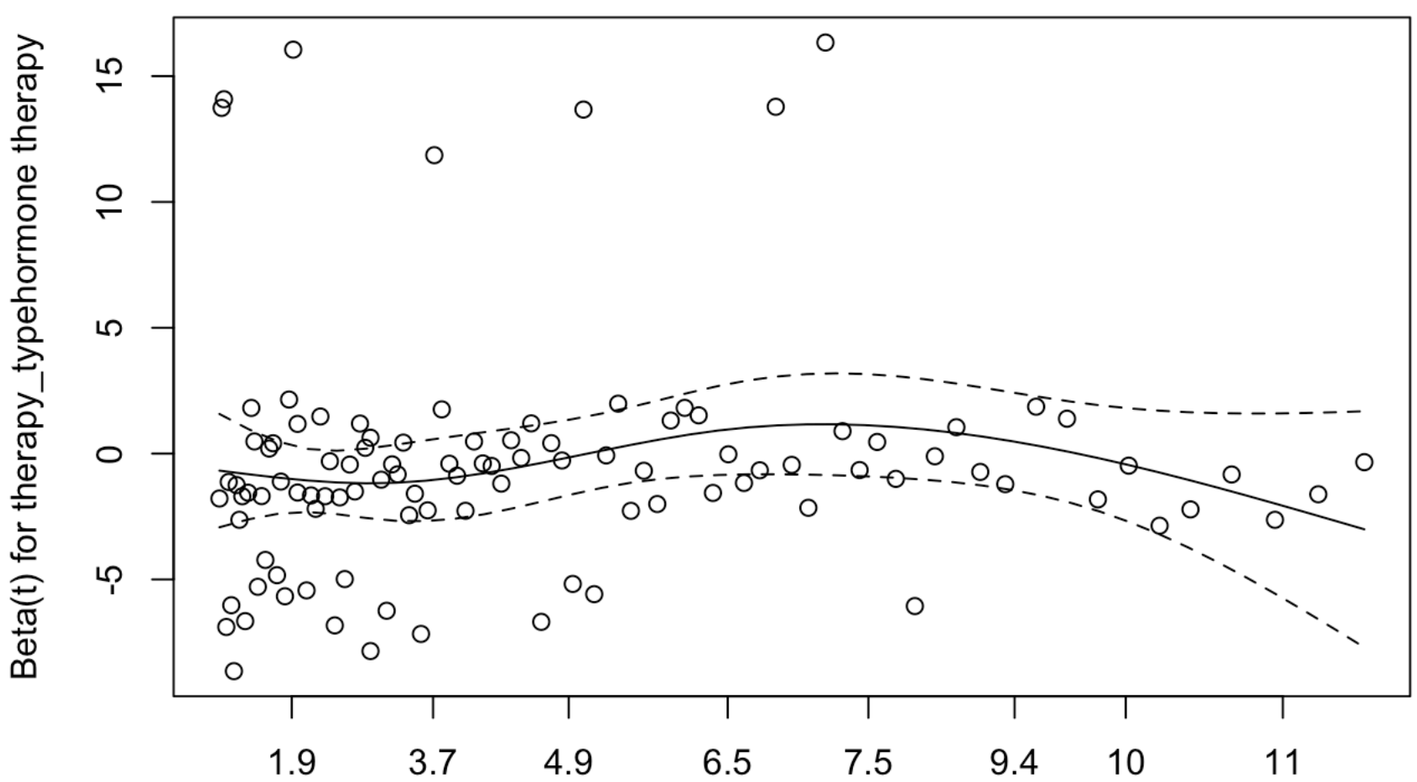
Cancer_stage is jointly determined by the tumor (T) stage, metastasis (M) stage, and lymph node (N) stage. So, in any linear model, we cannot incldue the overall cancer stage and the individual T,M,N stages as this will result in multicollinearity. So, we consider two separate models, one with only cancer stage and the other with the individual T,M,N stages.

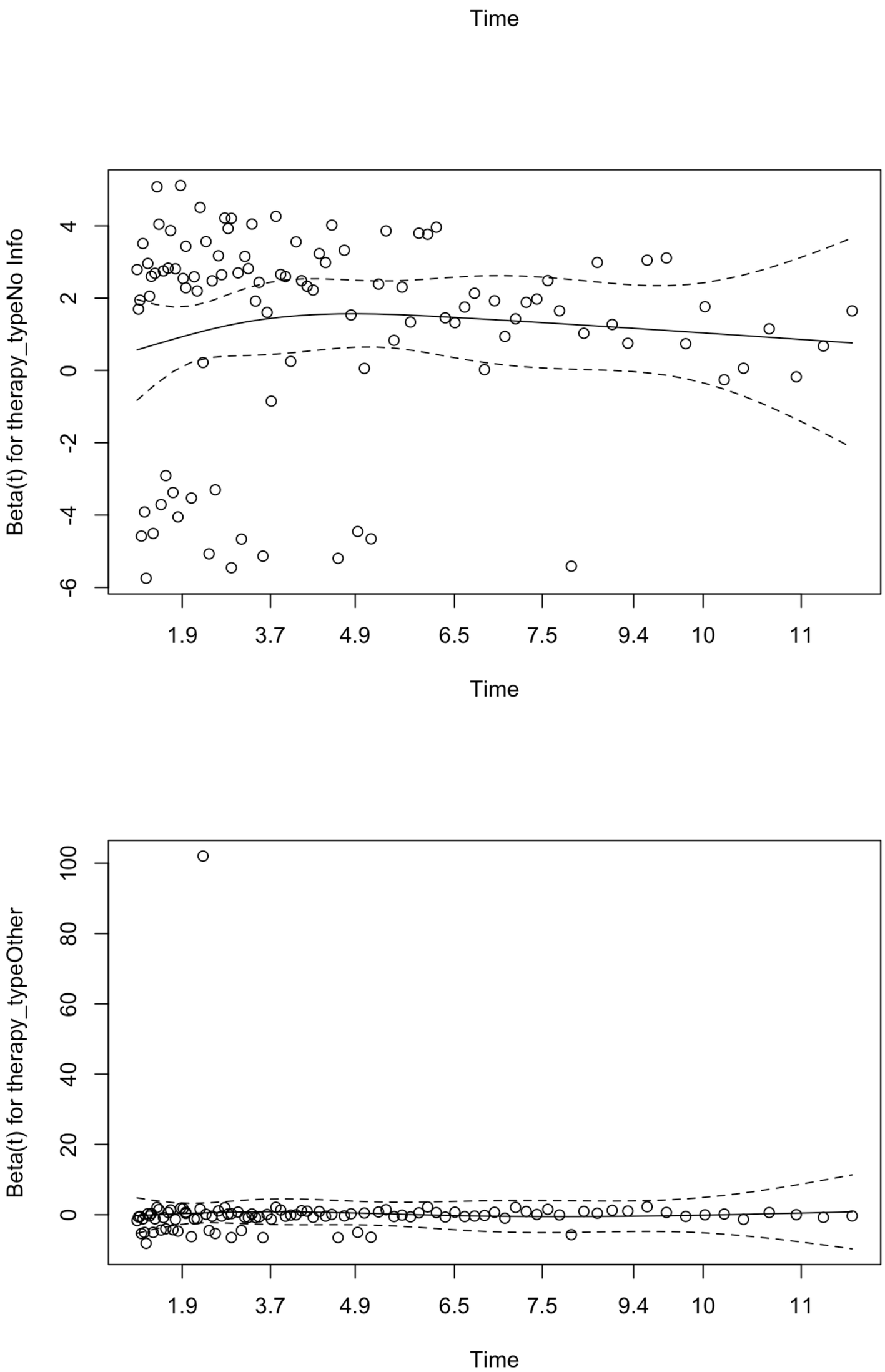
7.10.1 Cox model with Age, Therapy and Cancer Stages

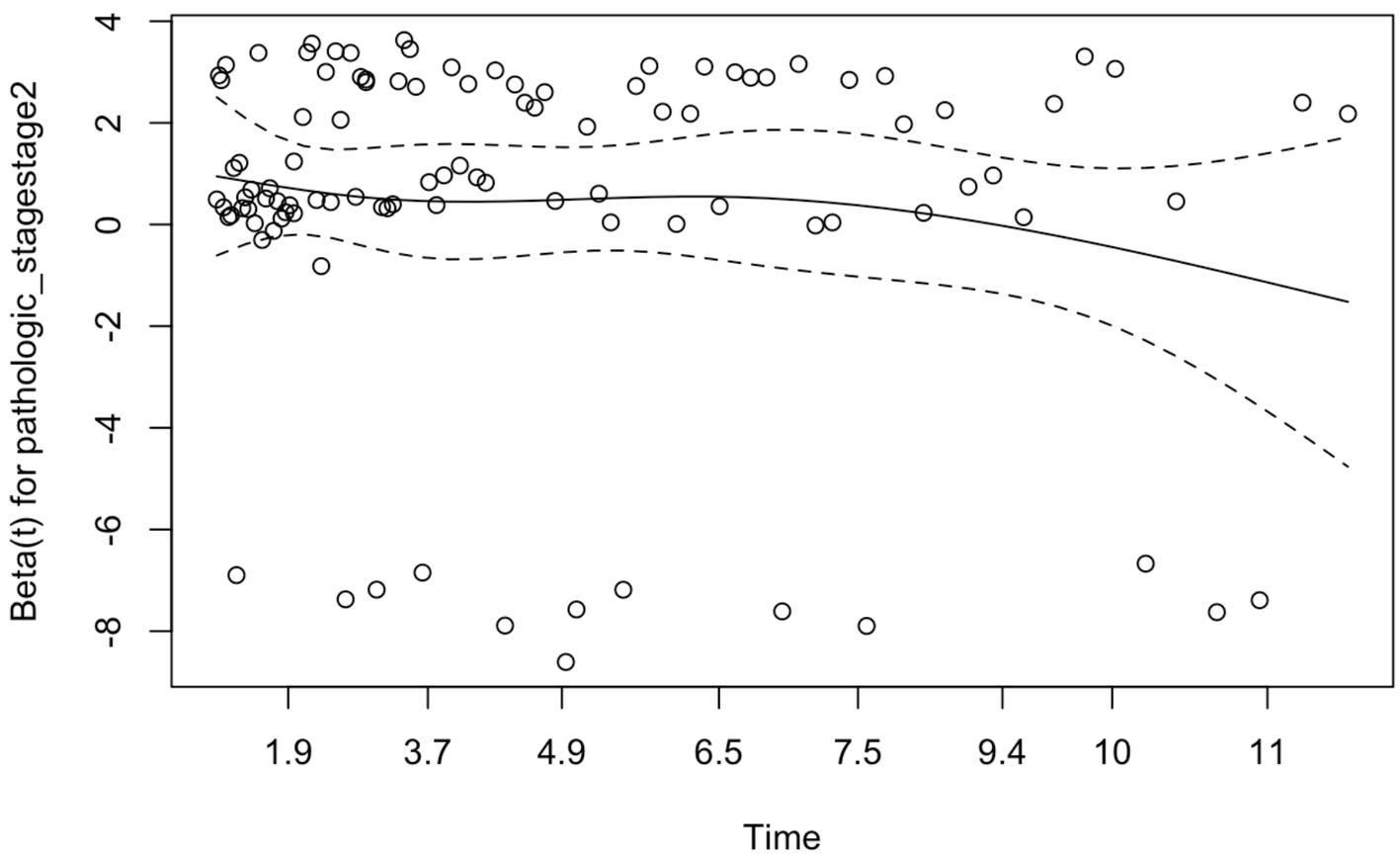
Check for Proportional Hazards Assumption

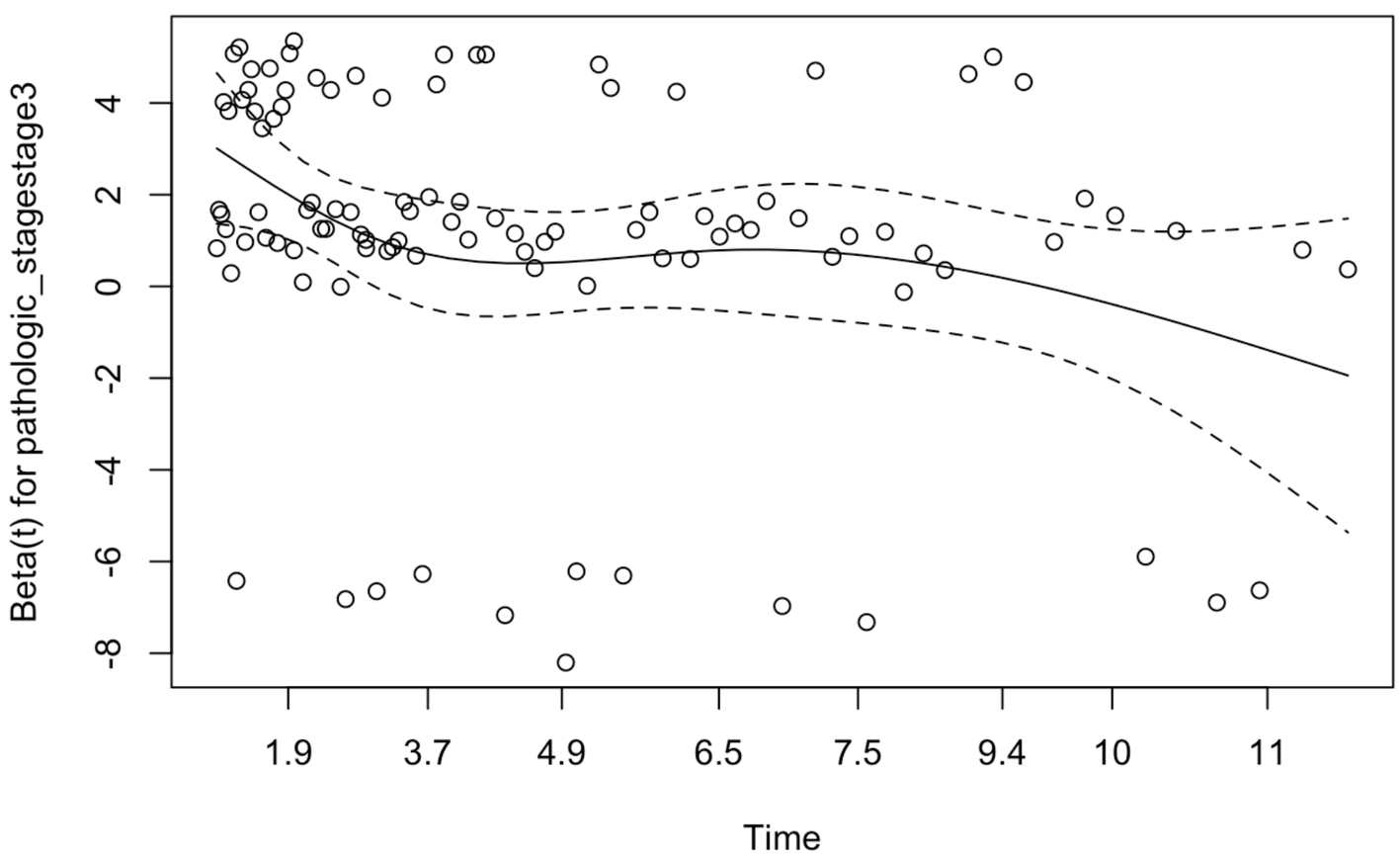
```
##                                 rho   chisq       p
## age                         -0.0584  0.4501 0.50229
## therapy_typehormone therapy  0.0642  0.4263 0.51381
## therapy_typeNo Info          0.0287  0.0825 0.77391
## therapy_typeOther           -0.0145  0.0210 0.88465
## pathologic_stagestage2      -0.1212  1.5160 0.21823
## pathologic_stagestage3      -0.2607  6.8801 0.00872
## pathologic_stagestage4      -0.2360  5.6315 0.01764
## pathologic_stagestageX      -0.1166  1.3832 0.23955
## GLOBAL                          NA 11.0924 0.19652
```

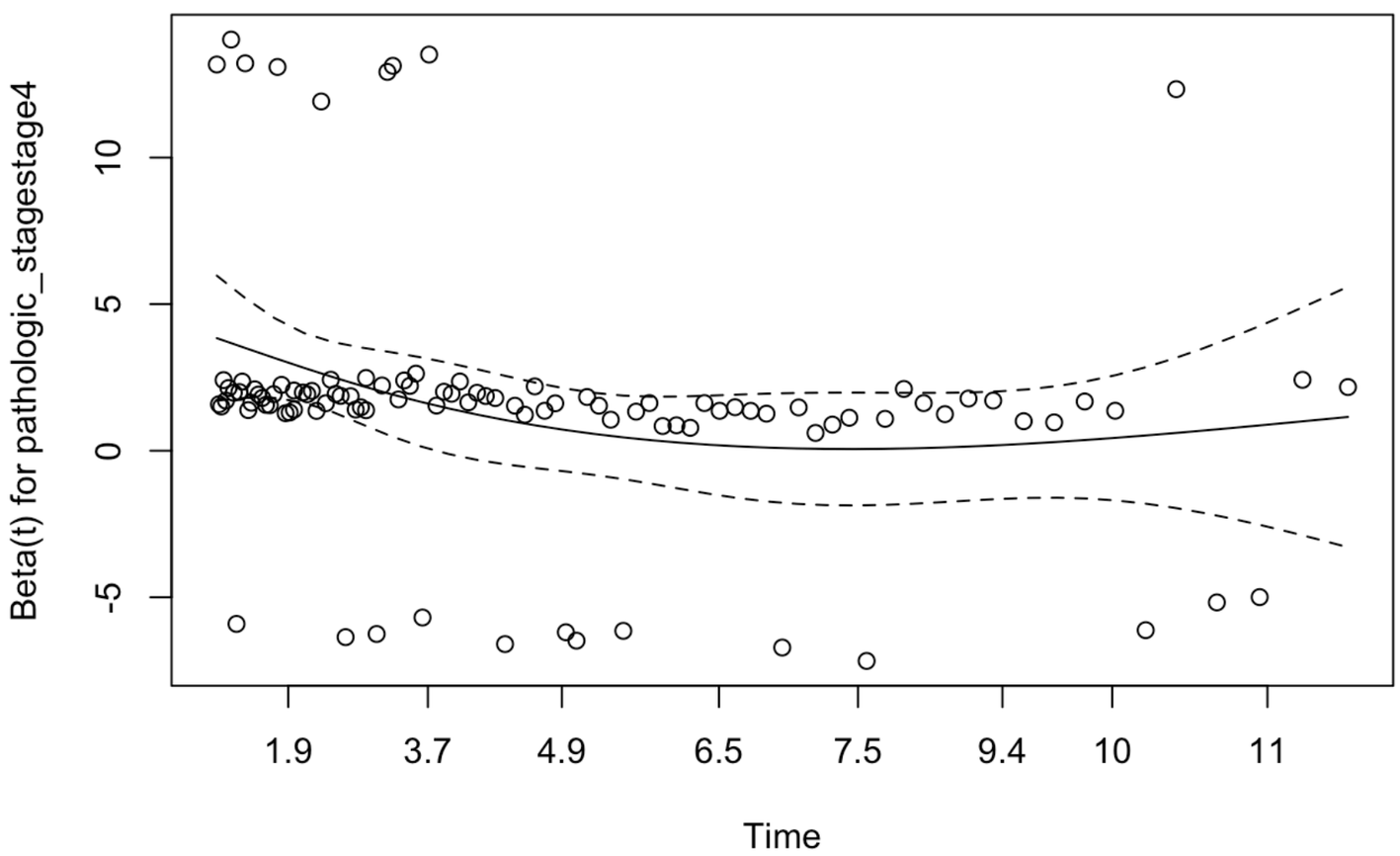

Beta(t) for age
0.15
0.05
-0.05
-0.15
1.9
3.7
4.9
6.5
7.5
9.4
10
11
Time

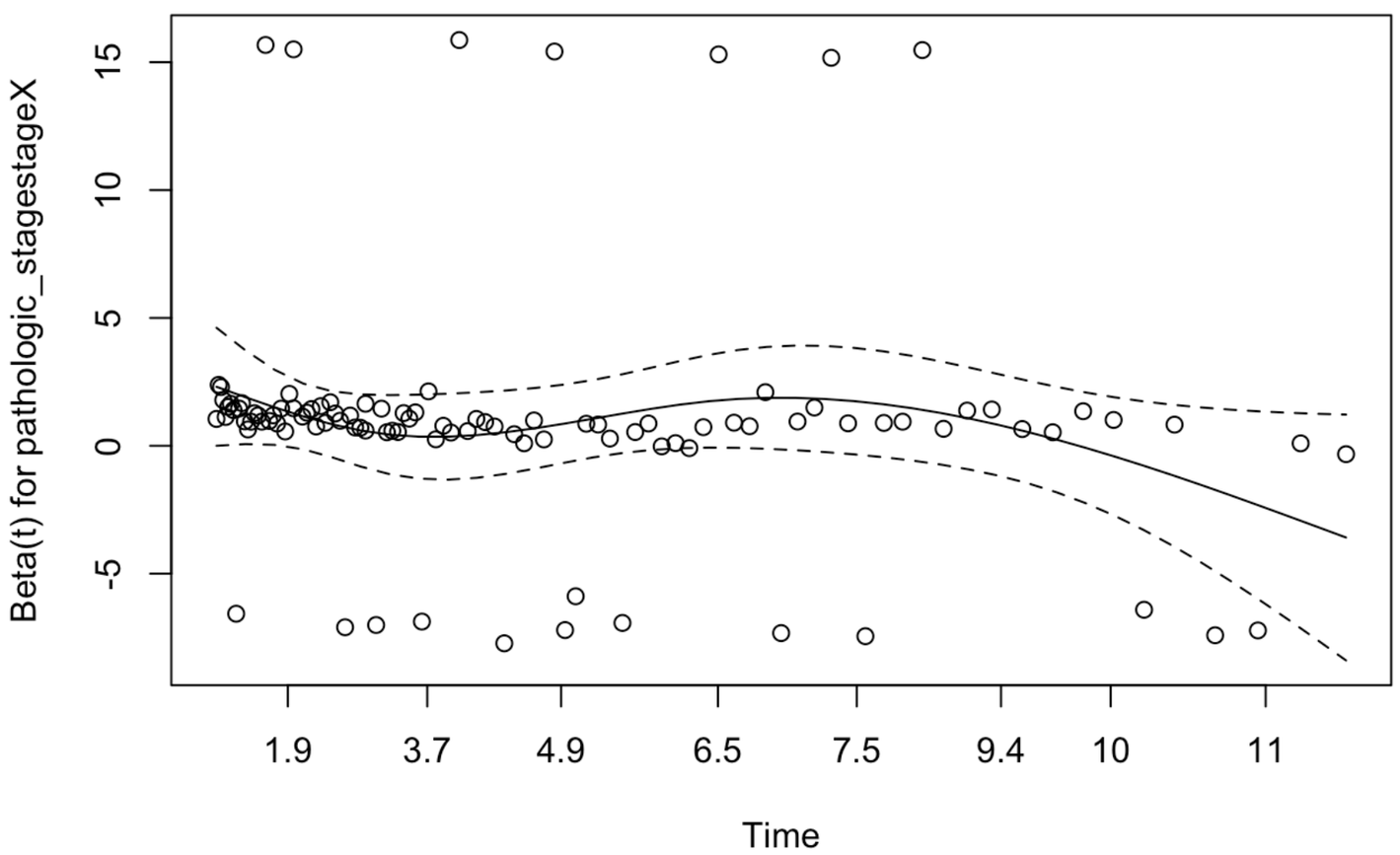

Beta(t) for therapy_typehormone therapy
15
10
5
0
-5
1.9
3.7
4.9
6.5
7.5
9.4
10
11


Time
Beta(t) for therapy_typeNo Info
-6
-4
-2
0
2
4
1.9
3.7
4.9
6.5
7.5
9.4
10
11
Beta(t) for therapy_typeOther
0
20
40
60
80
100
Time


Beta(t) for pathologic_stagestage2
4
2
0
-2
-4
-6
-8
1.9
3.7
4.9
6.5
7.5
9.4
10
11
Time

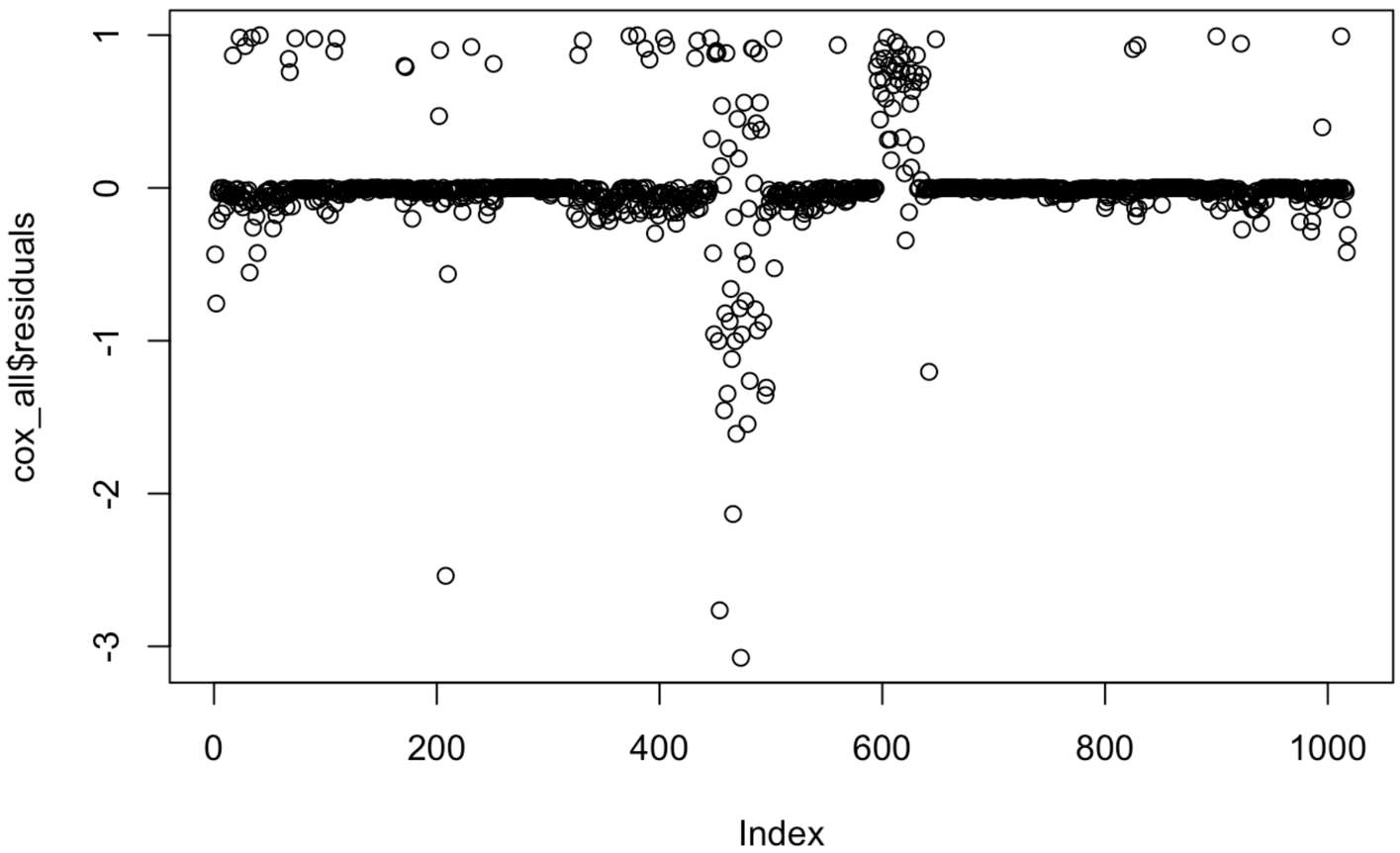

Beta(t) for pathologic_stagestage3
4
2
0
-2
-4
-6
-8
1.9
3.7
4.9
6.5
7.5
9.4
10
11
Time


Beta(t) for pathologic_stagestage4
10
5
0
-5
1.9
3.7
4.9
6.5
7.5
9.4
10
11
Time



The Schonfeld residuals and the p-value for the corresponding chi-square distribution are checked to verify the assumption of proportional hazards of the Cox model. The plot of the Schonfeld residuals shows a random distribution around a mean of 0. The corresponding p-values for the chi-square distribution are high indicating non-significance. This indicates that the proportional hazards condition is met.

```
## Call:
## coxph(formula = surv_obj_all ~ age + therapy_type + pathologic_stage,
##     data = brca_clin)
##
##   n= 1018, number of events= 103
##    (14 observations deleted due to missingness)
##
##                                coef exp(coef)  se(coef)      z Pr(>|z|)
## age                         0.023335  1.023610  0.007635  3.056  0.00224
## therapy_typehormone therapy -0.486062  0.615044  0.460042 -1.057  0.29071
## therapy_typeNo Info          1.178920  3.250861  0.284649  4.142 3.45e-05
## therapy_typeOther            0.224043  1.251125  1.032249  0.217  0.82817
## pathologic_stagestage2       0.453316  1.573521  0.318093  1.425  0.15413
## pathologic_stagestage3       1.093467  2.984604  0.335582  3.258  0.00112
## pathologic_stagestage4       1.685171  5.393373  0.435835  3.867  0.00011
## pathologic_stagestageX       0.972513  2.644582  0.471317  2.063  0.03908
##
## age                         **
## therapy_typehormone therapy
## therapy_typeNo Info         ***
## therapy_typeOther
## pathologic_stagestage2
## pathologic_stagestage3      **
## pathologic_stagestage4      ***
## pathologic_stagestageX      *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                             exp(coef) exp(-coef) lower .95 upper .95
## age                             1.024     0.9769    1.0084     1.039
## therapy_typehormone therapy     0.615     1.6259    0.2496     1.515
## therapy_typeNo Info             3.251     0.3076    1.8608     5.679
## therapy_typeOther               1.251     0.7993    0.1654     9.461
## pathologic_stagestage2          1.574     0.6355    0.8436     2.935
## pathologic_stagestage3          2.985     0.3351    1.5461     5.761
## pathologic_stagestage4          5.393     0.1854    2.2955    12.672
## pathologic_stagestageX          2.645     0.3781    1.0500     6.661
##
## Concordance= 0.776  (se = 0.029 )
## Likelihood ratio test= 73.6  on 8 df,   p=9e-13
## Wald test            = 70.29  on 8 df,   p=4e-12
## Score (logrank) test = 81.41  on 8 df,   p=3e-14
```



The base hazard model corresponds to age=0, treatment type of chemotherapy, and cancer stage 1. Of the other predictors, the coefficients for age, therapy_type of “No info”, cancer stages of 3,4, and X are found to be significant.

The Cox model equation can then be written as:

$$H(t|age, chemo, hormone, no_info, other, stage1, stage2, stage3, stage4, stageX) = H_0(t)exp(0.023 * age + 1.18 * No\_info + 1.09 * stage3 + 1.685 * stage4 + 0.97$$

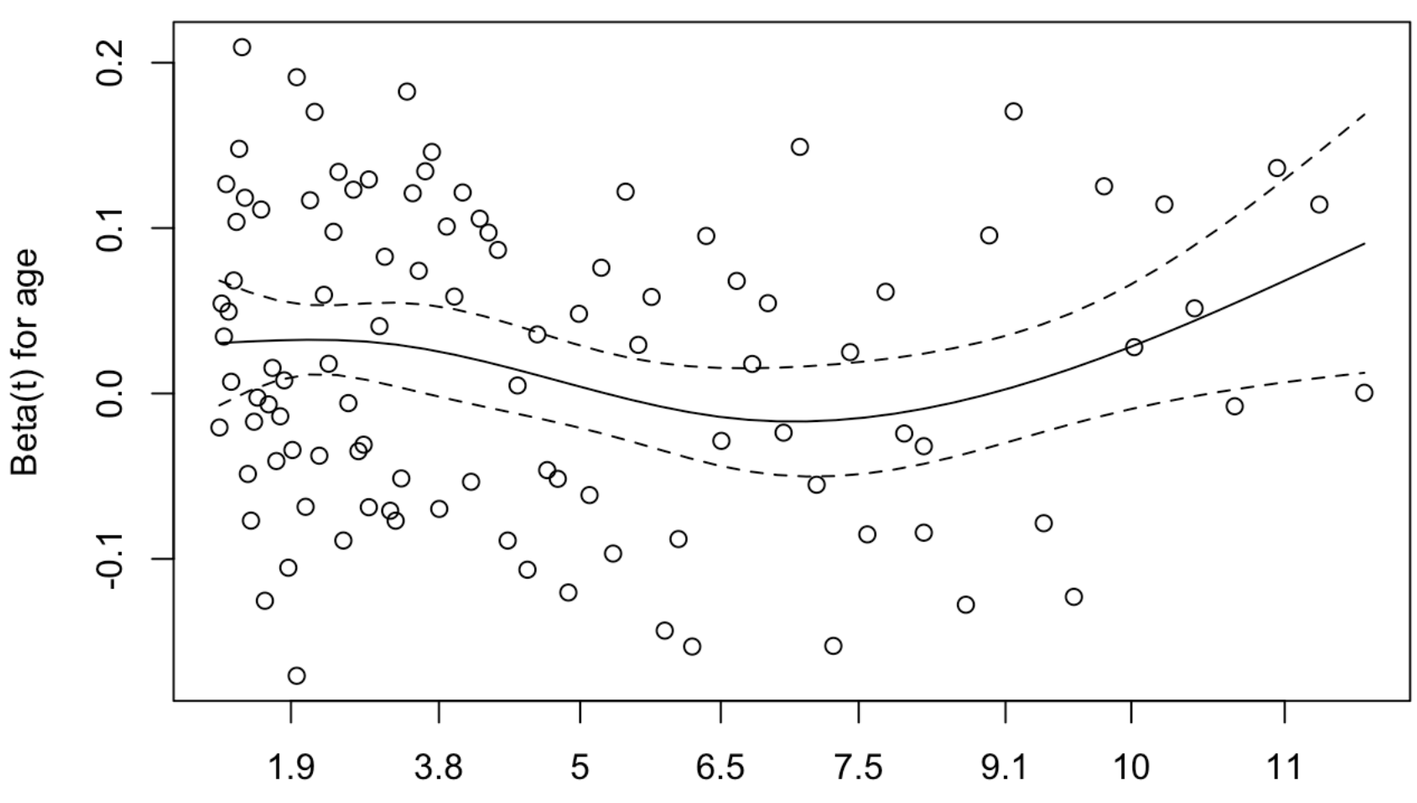
This equation can be used to compute the cumulative hazard rate for new values.

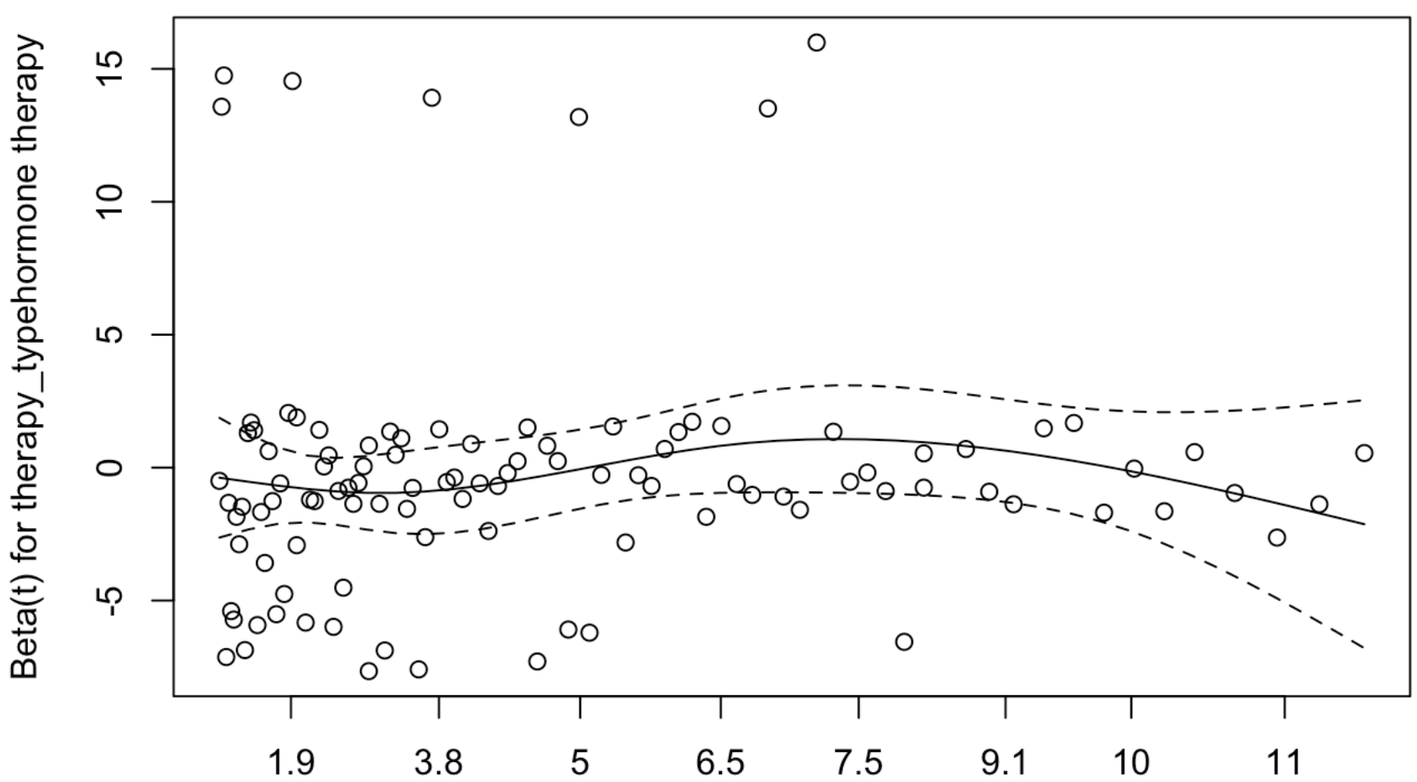
Considering only the signifcant predictors, the hazard ratio (exp(coef)) values are used to summarise the following findings. Recall that while analysing each of the predictors, all the other predictors in the model are accounted for by using their mean values.
1. For every one additional year of age, the risk of death increases by a factor of 1.02. 2. Compared to subjects on chemotherapy treatment, the risk of death for subjects with No info on the therapy increses by a factor 3.25 4. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 2.98 for subjects diagnosed with lymph node stage 3
5. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 5.39 for subjects diagnosed with lymph node stage 4
6. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 2.64 for subjects diagnosed with lymph node stage X (or not assessed) cancer

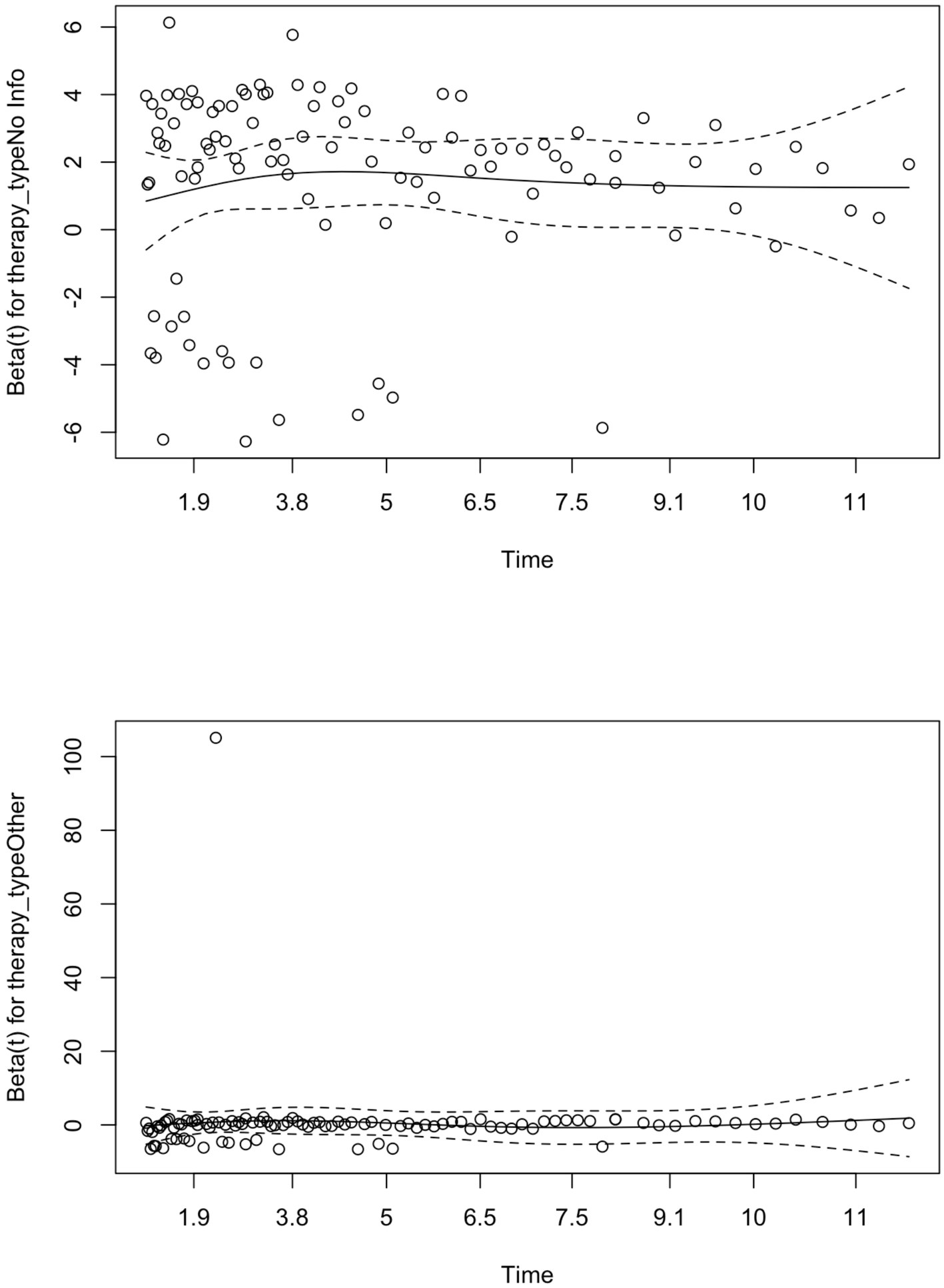
7.10.2 Cox model with Age, Therapy, Tumor, and Lymph Node Stages

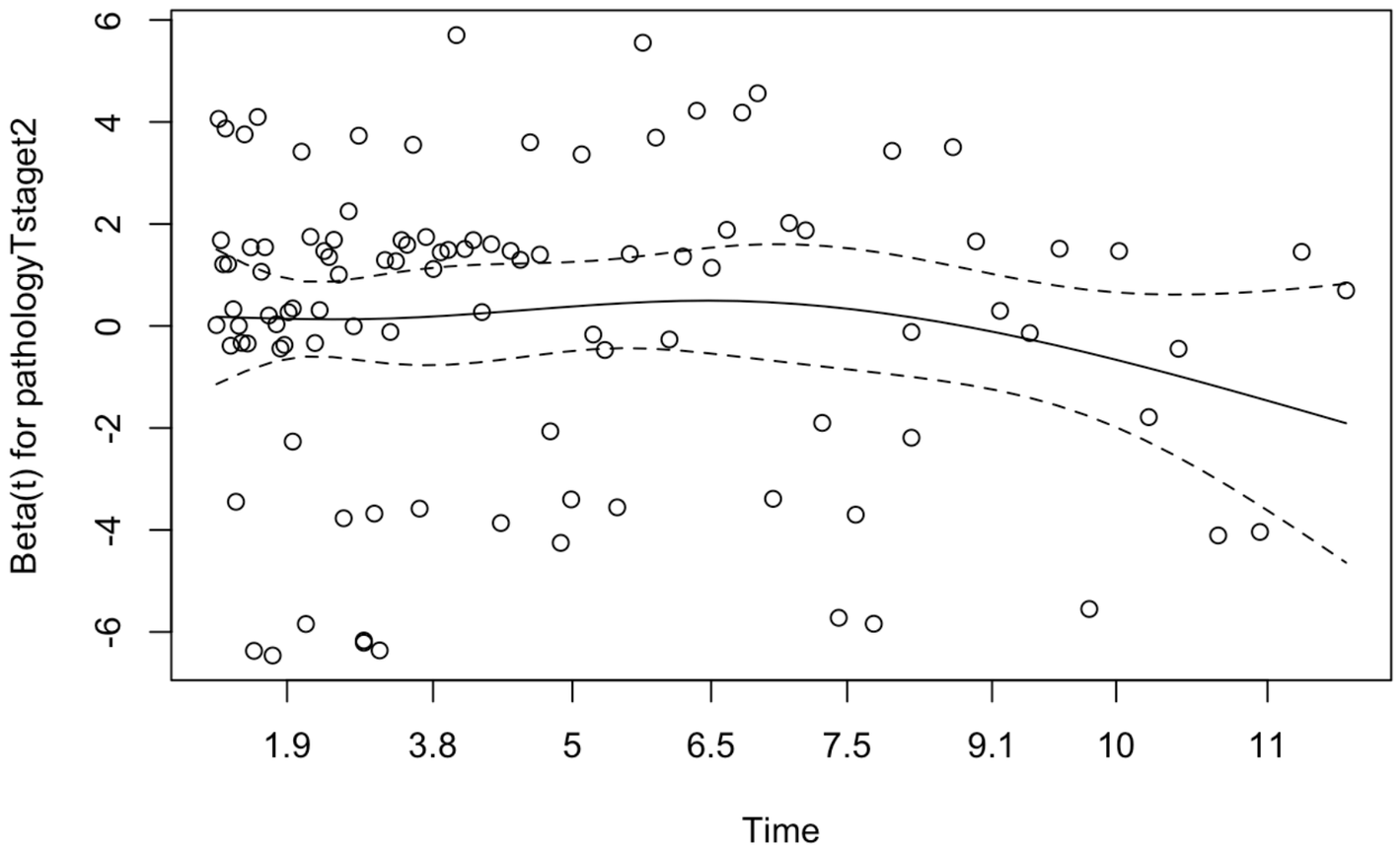
Check for Proportional Hazards Assumption

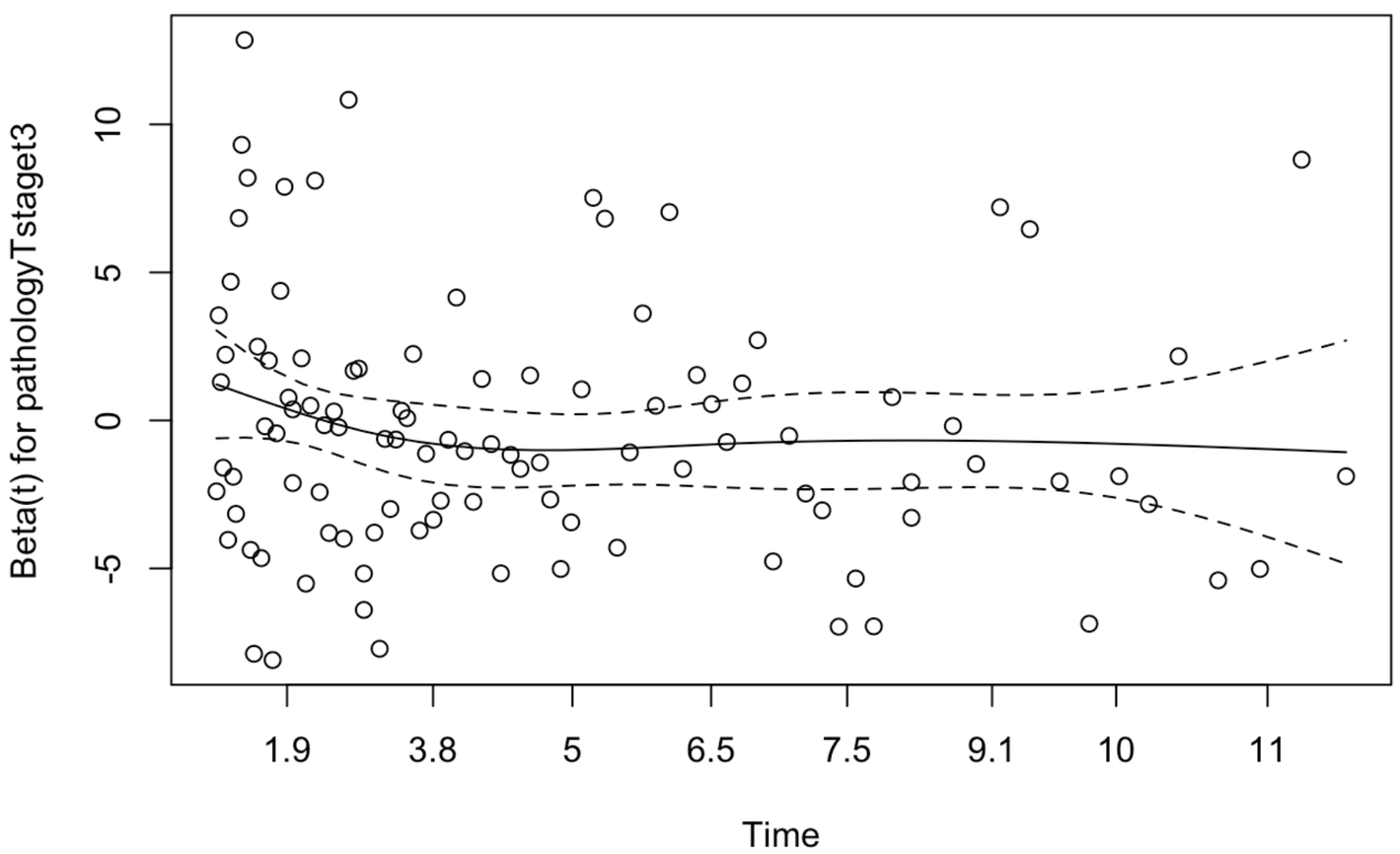
```
##                                rho  chisq      p
## age                        -0.05009 0.3503 0.5540
## therapy_typehormone therapy  0.06280 0.4004 0.5269
## therapy_typeNo Info          0.01915 0.0343 0.8532
## therapy_typeOther           -0.01431 0.0216 0.8833
## pathologyTstaget2           -0.07857 0.7247 0.3946
## pathologyTstaget3           -0.11074 1.6532 0.1985
## pathologyTstaget4           -0.18697 3.7787 0.0519
## pathologyTstagetx           -0.02583 0.0663 0.7968
## pathologyNstagen1           -0.00862 0.0080 0.9287
## pathologyNstagen2           -0.03101 0.1122 0.7376
## pathologyNstagen3           -0.02324 0.0576 0.8104
## pathologyNstagenx            0.04654 0.2330 0.6293
## GLOBAL                            NA 8.3903 0.7539
```

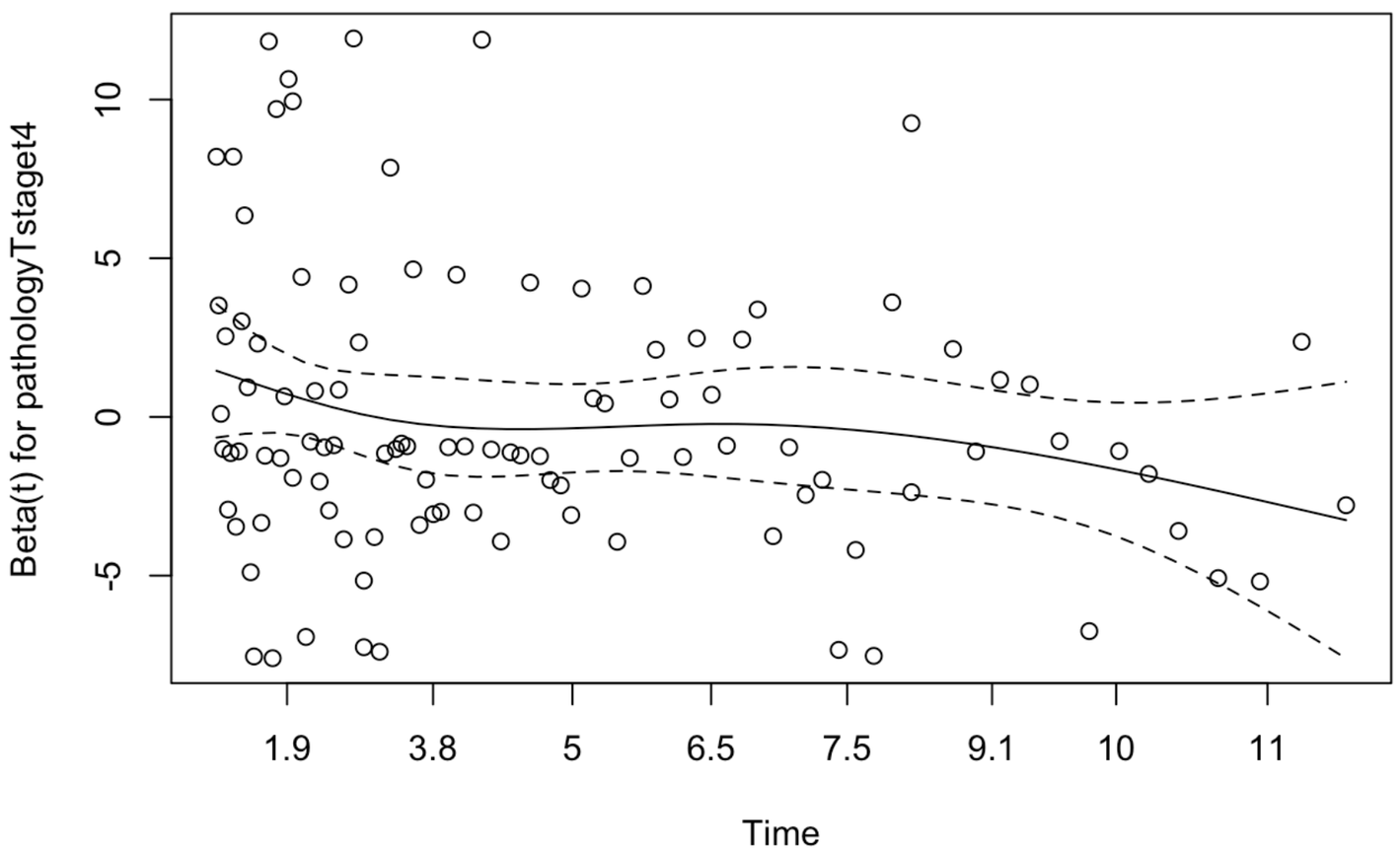

Beta(t) for age
0.2
0.1
0.0
-0.1
1.9
3.8
5
6.5
7.5
9.1
10
11
Time

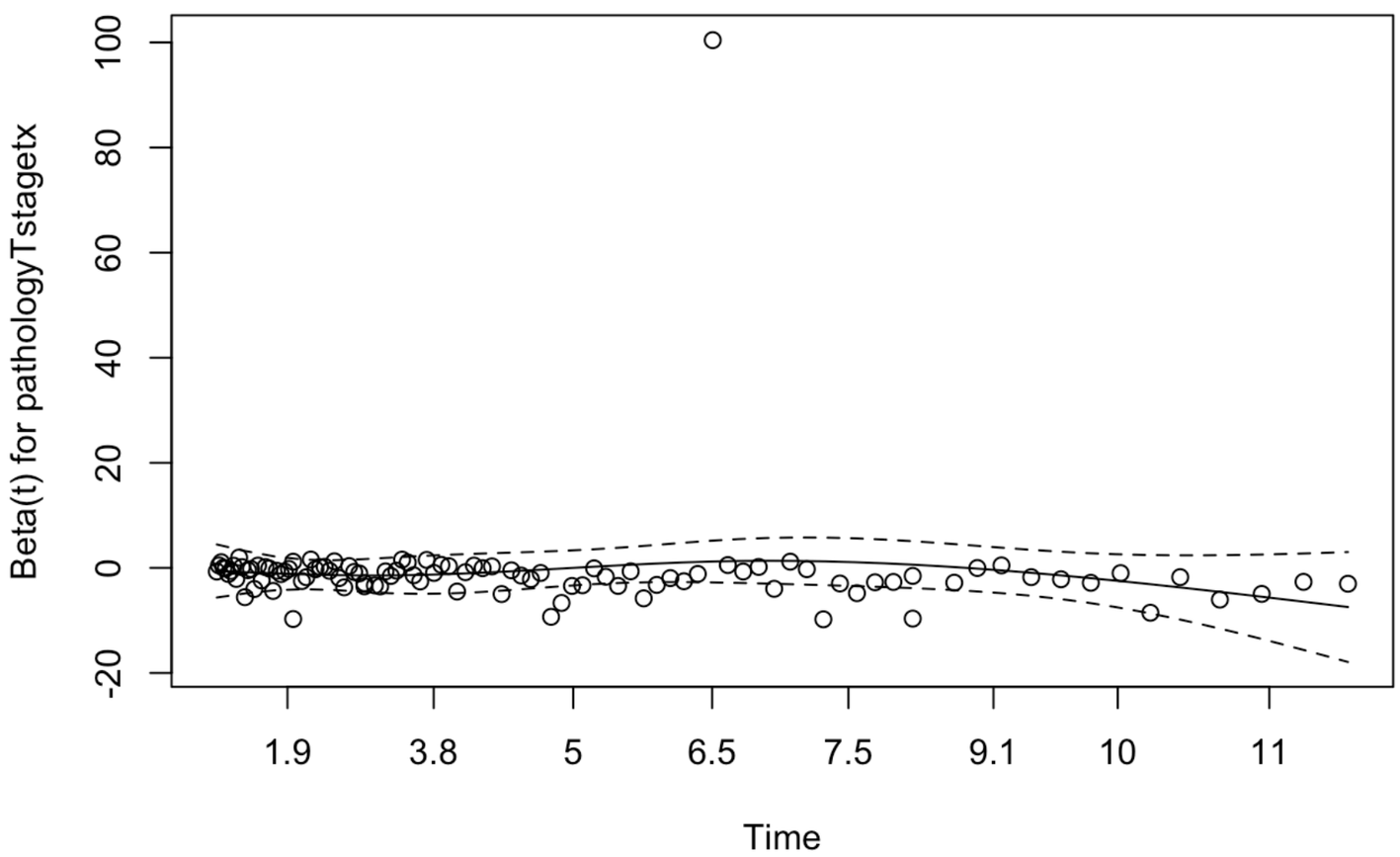

Beta(t) for therapy_typehormone therapy
15
10
5
0
-5
1.9
3.8
5
6.5
7.5
9.1
10
11

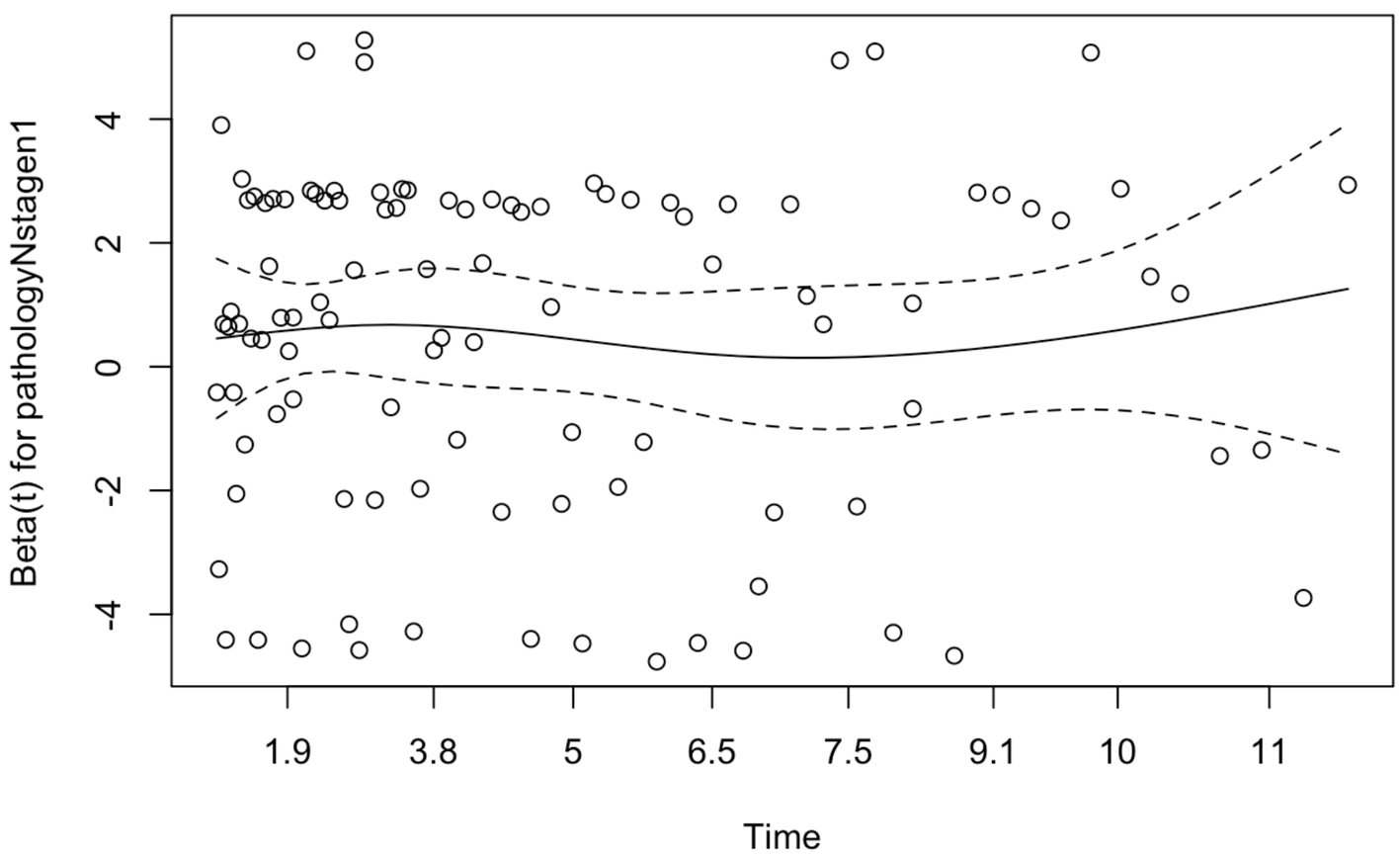

Time
Beta(t) for therapy_typeNo Info
6
4
2
0
-2
-4
-6
1.9
3.8
5
6.5
7.5
9.1
10
11
Beta(t) for therapy_typeOther
100
80
60
40
20
0
1.9
3.8
5
6.5
7.5
9.1
10
11
Time

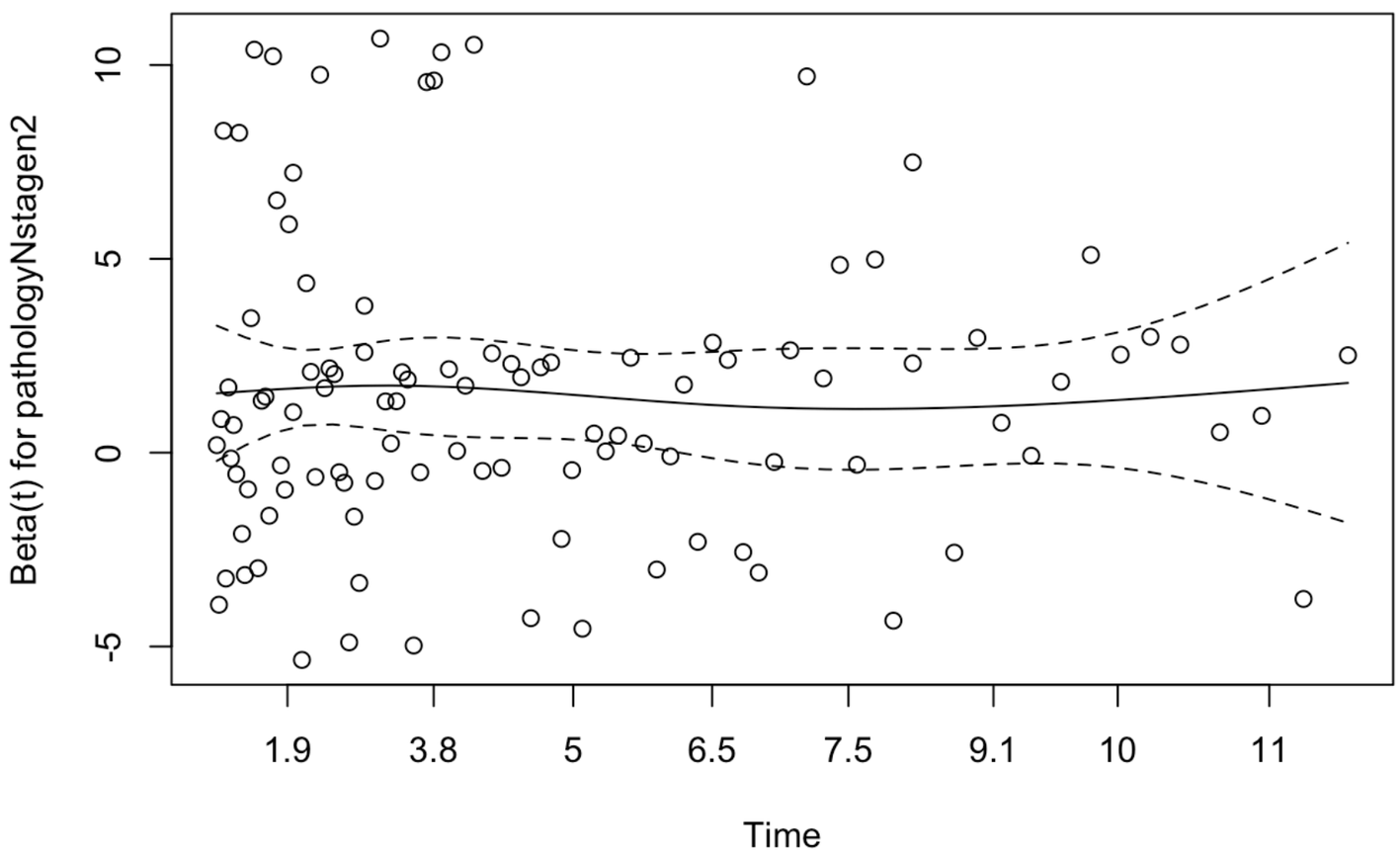

Beta(t) for pathologyTstaget2
6
4
2
0
-2
-4
-6
1.9
3.8
5
6.5
7.5
9.1
10
11
Time

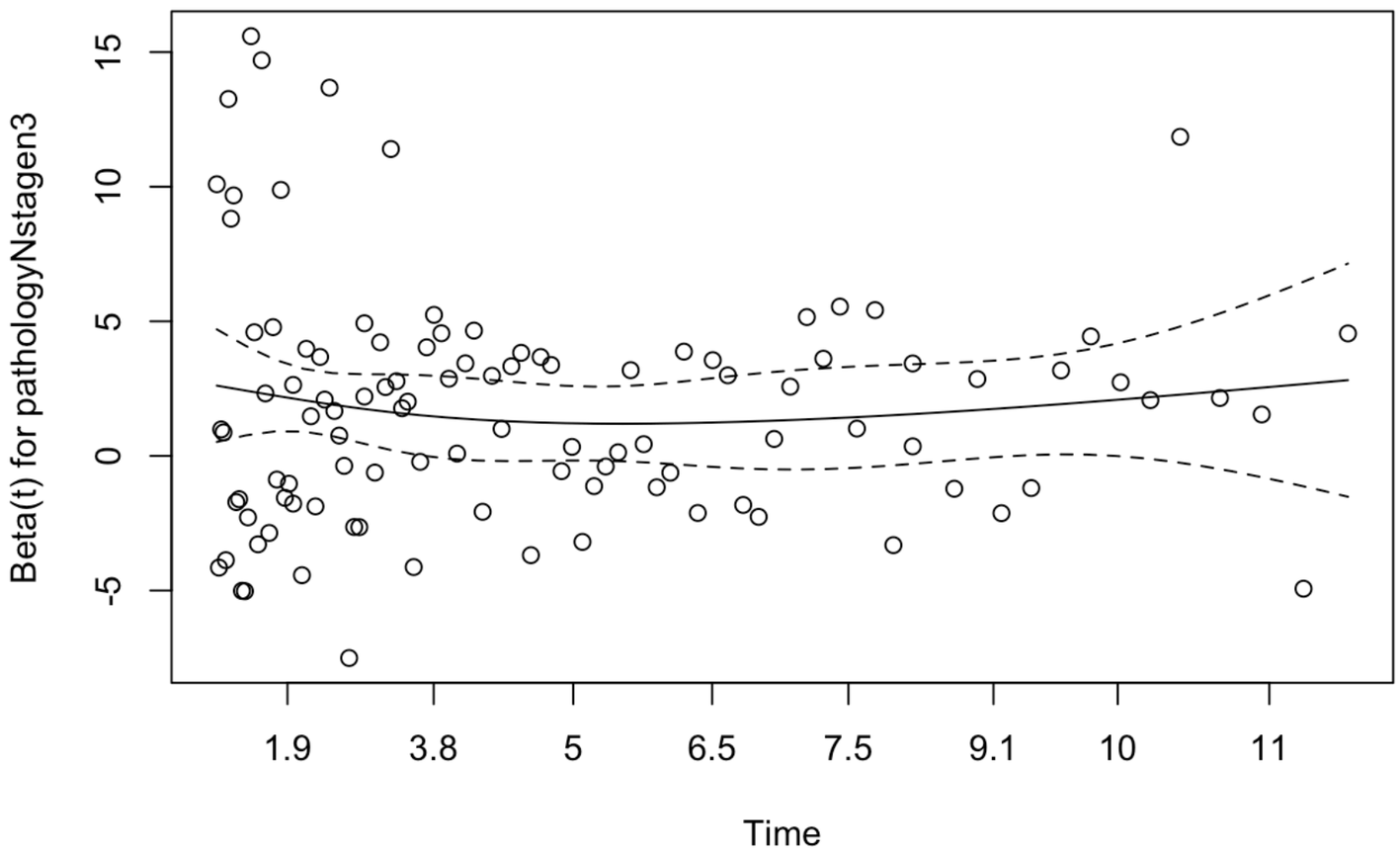

Beta(t) for pathologyTstaget3
10
5
0
-5
1.9
3.8
5
6.5
7.5
9.1
10
11
Time

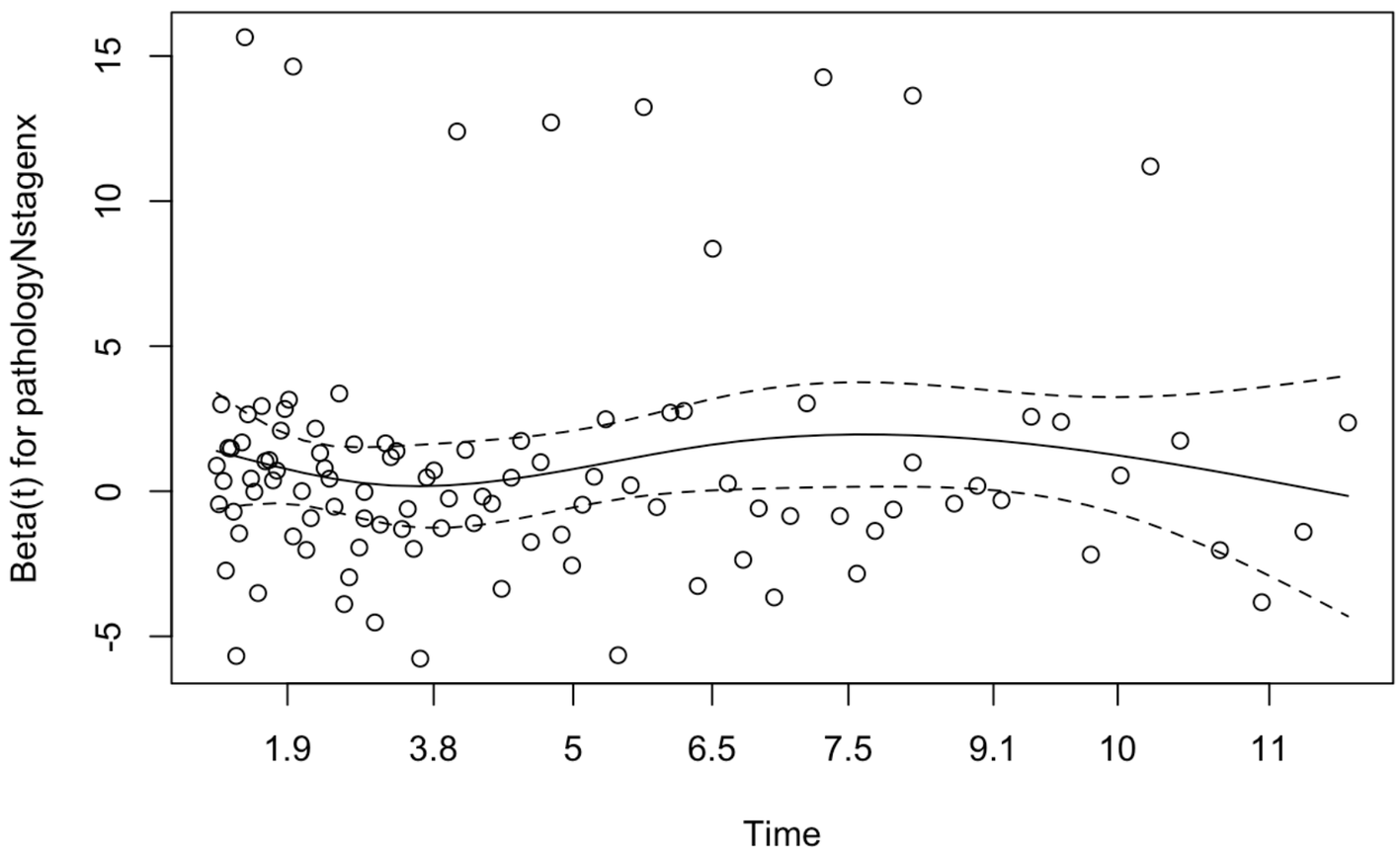

Beta(t) for pathologyTstaget4
10
5
0
-5
1.9
3.8
5
6.5
7.5
9.1
10
11
Time


Beta(t) for pathologyTstagetx
100
80
60
40
20
0
-20
1.9
3.8
5
6.5
7.5
9.1
10
11
Time


Beta(t) for pathologyNstagen1
4
2
0
-2
-4
1.9
3.8
5
6.5
7.5
9.1
10
11
Time


Beta(t) for pathologyNstagen2
10
5
0
-5
1.9
3.8
5
6.5
7.5
9.1
10
11
Time

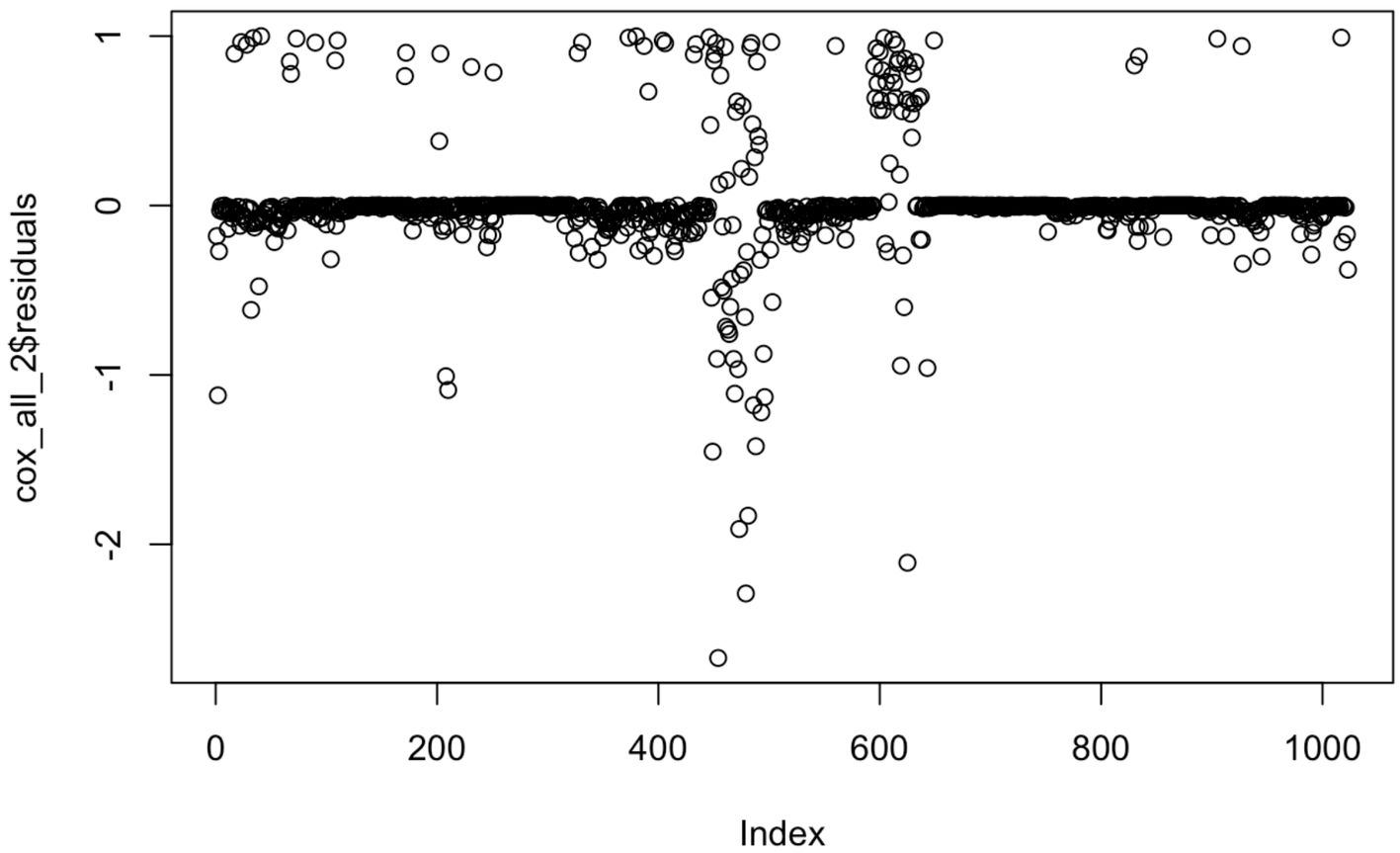




The Cox model is fit and verified that the hazards are proportional by checking the Schonfeld residuals. The plot of the Schonfeld residuals shows a random distibution around a mean of 0. The corresponding p-value for the chi-square distribution is high indicating non-significance. This indicates that the proportional hazards conditions are met.

Interpretation of the Model

```
## Call:
## coxph(formula = surv_obj_all ~ age + therapy_type + pathologyTstage +
##     pathologyNstage, data = brca_clin)
##
##   n= 1023, number of events= 104
##    (9 observations deleted due to missingness)
##
##                               coef exp(coef)  se(coef)      z Pr(>|z|)
## age                        0.019528  1.019720  0.007698  2.537   0.0112
## therapy_typehormone therapy -0.311487  0.732357  0.459920 -0.677   0.4982
## therapy_typeNo Info         1.378160  3.967596  0.294427  4.681 2.86e-06
## therapy_typeOther           0.385762  1.470735  1.030526  0.374   0.7082
## pathologyTstaget2           0.113103  1.119747  0.269611  0.420   0.6748
## pathologyTstaget3          -0.306975  0.735669  0.371933 -0.825   0.4092
## pathologyTstaget4          -0.053922  0.947506  0.429735 -0.125   0.9001
## pathologyTstagetx          -0.820902  0.440035  1.031807 -0.796   0.4263
## pathologyNstagen1           0.510979  1.666922  0.263314  1.941   0.0523
## pathologyNstagen2           1.521524  4.579199  0.356199  4.272 1.94e-05
## pathologyNstagen3           1.795731  6.023878  0.427133  4.204 2.62e-05
## pathologyNstagenx           0.908249  2.479976  0.409037  2.220   0.0264
##
## age                        *
## therapy_typehormone therapy
## therapy_typeNo Info        ***
## therapy_typeOther
## pathologyTstaget2
## pathologyTstaget3
## pathologyTstaget4
## pathologyTstagetx
## pathologyNstagen1          .
## pathologyNstagen2          ***
## pathologyNstagen3          ***
## pathologyNstagenx          *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                            exp(coef) exp(-coef) lower .95 upper .95
## age                           1.0197     0.9807   1.00445     1.035
## therapy_typehormone therapy    0.7324     1.3655   0.29733     1.804
## therapy_typeNo Info            3.9676     0.2520   2.22798     7.066
## therapy_typeOther              1.4707     0.6799   0.19514    11.085
## pathologyTstaget2              1.1197     0.8931   0.66013     1.899
## pathologyTstaget3              0.7357     1.3593   0.35489     1.525
## pathologyTstaget4              0.9475     1.0554   0.40812     2.200
## pathologyTstagetx              0.4400     2.2725   0.05824     3.325
## pathologyNstagen1              1.6669     0.5999   0.99491     2.793
## pathologyNstagen2              4.5792     0.2184   2.27820     9.204
## pathologyNstagen3              6.0239     0.1660   2.60795    13.914
## pathologyNstagenx              2.4800     0.4032   1.11244     5.529
##
## Concordance= 0.77  (se = 0.028 )
## Likelihood ratio test= 83.1  on 12 df,   p=1e-12
## Wald test            = 77.68  on 12 df,   p=1e-11
## Score (logrank) test = 90.62  on 12 df,   p=4e-14
```



The base hazard model corresponds to age=0, treatment type of chemotherapy, tumor stage 1 and lymph node stage 0 (n0 - no cancer in lymph nodes). Of the other predictors, the coefficients for age, therapy_type of “No info”, lymph node stages 3,4, and X are found to be significant.

The Cox model equation can then be written as:

$$H(t|age, chemo, hormone, no_info, other, t1, t2, t3, t4, tX, n0, n1, n2, n3, nX) = H_0(t)exp(0.02 * age + 1.38 * No\_info + 1.52 * n2 + 1.80 * n3 + 0.91 * stageX)$$

This equation can be used to compute the cumulative hazard rate for new values.

Considering only the signifcant predictors, the hazard ratio (exp(coef)) values are used to summarise the following findings. Recall that while analysing ech of the predictors, all the other predictors in the model are accounted for by using their mean values.
1. For every one additional year of age, the risk of death increases by a factor of 1.02. 2. Compared to subjects on chemotherapy treatment, the risk of death for subjects with No info on the therapy increses by a factor of 3.96 4. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 4.6 for subjects diagnosed with stage n2 5. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 5.9 for subjects diagnosed with stage n3 6. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 2.5 for subjects diagnosed with stage nX (or not assessed) cancer

7.10.2.1 Cox Proportional Hazard Model for All - with interactions

THe final step is to build a Cox PH model considering the interactions between each of the the variables - two-way, three-way, and four-way where applicable. When this is attempted, the Cox does not converge when interactions are considered between any sets of variables. So, this indicates that the predictors are independent.

8 Summary of Results

The following findings can be summarised:

1. No predictors:
    a. Kaplan Meir Results: Median Survival Time: 9.5 years
2. Age at first diagnosis:
    a. Kaplan Meir Results:
        a. For the young group (age < 40), the initial risk rate is high.
        b. Survival curves for the middle (40-60) and old age (>60) groups differe significantly.
        c. Median survival time is 8.4, 12.2, and 9.4 years for the young, middle and old age groups.
    b. Cox Proportional Hazards Model:
        a. Age is used as a continuous variable.
        b. The risk of death increases slowly by a factor 1.02 for every increase in age by a year.
    c. Parametric Model:
        a. Weibull distribution fits the survival data.
        b. When age is used as acontinuous varaible, it is found that The risk of death increases slowly by a factor 1.02 for every increase in age by a year.
        c. When age is used as a categorical varialbe (young, middle, old), it is found that the risk of death for the middle age group (ages 40-60) is almost 45% less (HR=0.55) than the risk of death for the younger group (age < 40) and the risk of death for the older age group (age > 60) is slightly higher by 10% compared to the younger age group. Equivalently, the survival time for the middle age group is higher by 46 % (ETR=1.46) for the middle age group when compared with the younger group and the survival time for the older age group is less by 95% when compared with the younger age group.
3. Race:
    a. Kaplan Meir Results: No difference in the survival curves between the races
    b. Parametric Model: Models are not good fits
    c. Cox Proportional Hazards Model: No significant difference between the races.
4. Ethnicity:
    a. Kaplan Meir Results: Median age for non-hispanc group is 9.51 years. No sufficient data for other group
    b. Parametric Model: Not good models

    c. Cox Proportional Hazards Model: Not a good fit.
5. Therapy Type:
    a. Kaplan Meir Results: “No Info” group is significantly different from the other therapy groups witht he survival of probability descreasing rapidly with no treatment.
    b. Parametric Model: Only exponential model fits. The constant hazard rate is 0.04 events/year and risk of death increases by 4.87 with no treatment vs chemotherapy
    c. Cox Proportional Hazards Model: The risk of death increases by a factor of 4 with no treatmentvs chemotherapy.
6. Cancer Stage Diagnosis:
    a. Kaplan Meir Results: Median survival time decreases with increase in diagnosis of cancer stage number from 10.8 years to 3.7 years. The survival curves of all stges are different from each other except between stage 1 and stage 2.
    b. Parametric Model:
        i. Weibull is the better model.
        ii. From the converted results of the weibull model, the risk of death increases with increase in the cancer stage number. When compared with the risk of death for stage 1, the risk is higher by a factor of 1.3 for stage 2, by a factor of 2.5 for stage 3 and by a factor of 6.6 for stage 4. Equivalently, the survival time for the stage 2 group reduces by 16% for stage 2, by 45% for stage 3 and by 70% for stage 4 when comapred with the survival time length for stage 1.
    c. Cox Proportional Hazards Model: Proportional hazards condition not met. Splitting up the datasets on time does not yield good Cox models.
7. Tumor Stage:
    a. Kaplan Meir Results: Median survival time decreases with increase in stage number of the tumor from 10 years for the first three stages to 4.5 years for stage t4. The survival curves are not found to be significantly different.
    b. Parametric Model:
    c. Cox Proportional Hazards Model: Proportional Hazards condition is not met. Splitting up the datasets on time does not yield good Cox models.
8. Metastasis Stage:
    a. Kaplan Meir Results: The median survival time is 10.5 years for stage m0 , with no metastasis. The survival curves for stages m0 and m1 are just slightly different from each not but statistically significant.
    b. Parametric Model: Both models are not good fits.
    c. Cox Proportional Hazards Model: Cox model is not a good fit as it results in infinite values.
9. Lymphnodes Stage:
    a. Kaplan Meir Results: Survival time reduces exponentially with increase in the number of lymph nodes with metastasis. The median survival time reduces from 11.7 years to 6.7 years. Survival curve for stage n0 (no metastasis in lymph nodes) is significantly different from survival curces
    b. Parametric Model: Weibull is the best fit model. when compared to the group of subjects with no metastasis in the lymph nodes, the risk of death increases by a facgor of 1.98 when the metastasis increases to 1-3 lymph nodes, increases by a factor of 3.3 when the metastasis increases to 4-9 lymph nodes and by a factor of 5.5 whe metastasis increases to 10 or more lymph nodes.
    c. Cox Proportional Hazards Model: Cox model is a good fit and coefficients for all stages are significant. Hazard ratio and hence the risk of death increases for each stage relative to the stage with no metastasis. When compared to the group of subjects with no metastasis in the lymph nodes (stage n0), the risk of death increases by a factor of 1.95 when the metastasis increases to 1-3 lymph nodes (stage n1), increases by a factor of 3.15 when the metastasis increases to 4-9 lymph nodes (stage n2) and by a factor of 5.22 when metastasis increases to 10 or more lymph nodes (stage n3).
10. Combined:
    a. Cox Proportional Hazards Model (Age, Therapy, Cancer Stage): The base hazard model corresponds to age=0, treatment type of chemotherapy, and cancer stage 1. Of the other predictors, the coefficients for age, therapy_type of “No info”, cancer stages of 3,4, and X are found to be significant.
        i. For every one additional year of age, the risk of death increases by a factor of 1.02.
        ii. Compared to subjects on chemotherapy treatment, the risk of death for subjects with No info on the therapy increases by a factor 3.25
        iii. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 2.98 for subjects diagnosed with lymph node stage 3
        iv. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 5.39 for subjects diagnosed with lymph node stage 4
        v. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 2.64 for subjects diagnosed with lymph node stage X (or not assessed) cancer
    b. Cox Proportional Hazards Model (Age, Therapy, Lymph Nodes Stage): The base hazard model corresponds to age=0, treatment type of chemotherapy, tumor stage 1 and lymph node stage 0 (n0 - no cancer in lymph nodes). Of the other predictors, the coefficients for age, therapy_type of “No info”, lymph node stages 3,4, and X are found to be significant.
        i. For every one additional year of age, the risk of death increases by a factor of 1.02.
        ii. Compared to subjects on chemotherapy treatment, the risk of death for subjects with No info on the therapy increses by a factor of 3.96
        iii. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 4.6 for subjects diagnosed with stage n2
        iv. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 5.9 for subjects diagnosed with stage n3
        v. Compared to subjects diagnosed with stage 1 cancer, the risk of death increases by a factor of 2.5 for subjects diagnosed with stage nX (or not assessed) cancer

9 Answers to Research Questions:

1. What is the median survival time for breast cancer? In the absence of any predictors, the median survival time is 9.5 years.
2. How do different cancers (breast, ovarian, lung) affect survival rates? Of the three types of cancers BRCA (breast cancer), GBM (Glioblastoma Multiforme) and OV (Ovarian Cancer) that were compared, BRCA has the highest survival rate.
3. What are the important factors that influence risk of death for Breast Cancer? When considered individually or together age at first diagnosis, no therapy, cancer stage 3 and cancer stage 4, and any metastasis into the lymph nodes significantly reduce survival time.
4. What are the effects of each factor on survival?
   The number of lymph nodes that have been metasized into seems to be the strongest predictor of survival.
5. What is the probability of survival for breast cancer when other clinical covariates are considered? The results do not change when the predictors are all accounted for indicating an additive model with no interactions.

10 References