Overview of Survival Analysis and Analysis of Breast Cancer Patients in Remission

Umaimah Ahmed *
Department of Mathematics and Statistics, Amherst College

December 18, 2023

Abstract

This paper introduces the main concepts of survival analysis. Survival analysis is an area of statistics that is concerned with analyzing the time until an event of interest occurs. This paper goes over the concepts of survival and hazard functions, censored data, Kaplan-Meier curves, the log-rank test, and the Cox proportional hazards model. Then, it applies these concepts to a real life dataset from the German Breast Cancer Study Group.

Keywords: 3 to 6 keywords, kaplan-meier, cox, proportional, hazard, time-to-event

*The author gratefully acknowledges Professor Wagaman for her support and understanding, as well as the faculty of the Amherst College Statistics Department for their guidance over the past several years.

1 Introduction

Broadly, survival analysis refers to statistical methods used for data where the outcome of interest is time until a certain event occurs. This event could be a multitude of things such as death, bankruptcy, divorce, recovery, or relapse from remission, to name a few. Survival analysis can be applied to many different fields from medicine, to behavioral science, to business.

In this paper I will discuss the use of survival analysis and how it is affected by censoring. Then I will introduce and demonstrate the uses of the Kaplan-Meier curve and the Cox Proportional Hazards Model. After this, I will demonstrate how survival analysis can be performed on a real life dataset.

2 Survival Analysis

2.1 Survival and Hazard Functions

In survival analysis, survival and hazard probabilities are used to model data.

$S(t)$ is also often referred to as the survival function or survivor function and has the formula,

$S(t) = Pr(T > t)$.

The survival curve gives "the probability that an individual survives from the time origin to the specified future time t", where t = time until event occurs [Clark et al., 2003].

Another quantitative term essential to survival analysis is the hazard function, $h(t)$. The hazard function is given by the formula,

$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$.

The hazard is "the probability that an individual who is under observation at a time t

has an event at that time. Put another way, it represents the instantaneous event rate for an individual who has already survived to time t" [Clark et al., 2003].

It is important to note that the survival function gives the probability at t of the event NOT occurring while the hazard function gives the potential per unit of time for the event to occur.

2.2 Censoring

A problem that often comes up in survival analysis is censoring. Censoring occurs when we only have partial information on an individuals survival time.

In Kleinbaum and Klein's textbook, *Survival Analysis: A Self-Learning Text*, the authors list three reasons why censoring may occur:

1. a person does not experience the event before the study ends;
2. a person is lost to follow-up during the study period;
3. a person withdraws from the study because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction or other competing risk) [Kleinbaum and Klein, 2020, pp.6].

The visualization below shows an example from the text representing different ways censoring can occur.

```
knitr::include_graphics("gfx/censoring_fig.png")
```

Cases B, D, and E are examples of of data that is right-censored. For these data, we do not know the full survival time interval because the event did not occur before the end of the trial or the observation was lost to follow-up.

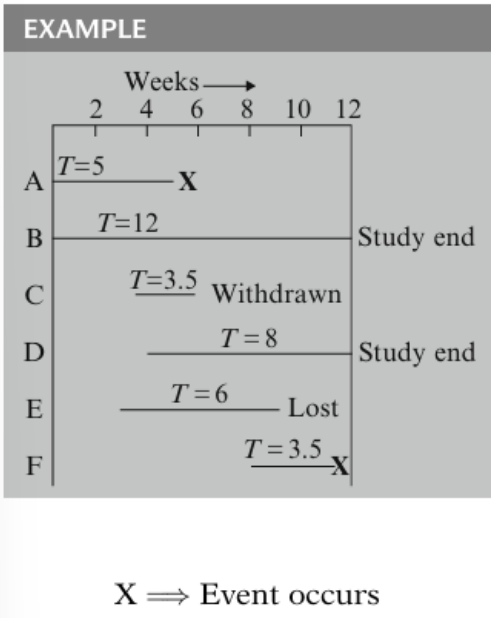


Figure 1: Examples of censored data from Survival Analysis: A Self-Learning Text

Data can also be left-censored. This is when the "true survival time is less than or equal to the observed survival time" [Kleinbaum and Klein, 2020, pp. 8]. For example, if the event of interest is contracting COVID-19, we can only record the event occurring after a patient has tested positive for the virus. The true survival time may have actually been shorter because we do not know the exact time the patient was first exposed to the virus.

2.3 Kaplan Meier method and Log Rank Test

We can use the Kaplan-Meier method, or the KM method, to graph survival curves and the log-rank test to test their equivalence. A KM curve graphically represents the survival function in a plot with time until an event on the x-axis and the survival probability on the y-axis.

We can use the log-rank test to compare two or more survival curves. When we compare two survival curves against each other, we test if they are equivalent or not. The log-rank test is a χ^2 test with 1 degree of freedom. We test with the null hypothesis that there is

no difference between the two survival curves and the alternate hypothesis that there is a significant difference in the survival curves. The formula for the test statistic is:

$\chi^2 = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2}$,

where O_1 and E_1 are the total number of observed and expected events, respectively, in group 1; and O_2 and E_2 are the total number of observed and expected events in group 2 [Bewick et al., 2004].

2.3.1 Examples in Literature

In a recent study, researchers performed a survival analysis of COVID-19 in the Mexican population. This study was done on data from Mexican Ministry of Health and included information on the survival time, age, sex, history of related illnesses, and ICU admission, among other things, of 16,752 COVID positive patients. Through surival analysis, the research team found that the risk of dying at any time was higher for men, older individuals and people with chronic kidney disease [Salinas-Escudero et al., 2020].

Below are some Kaplan-Meier curves presented in the study for factors that were found to be statistically significant. In Figure 1A, we can see that for nearly any point in time, survival probability drops drastically for the older age groups. In Figure 1C and 1D, we can see that having illnesses such as pneumonia and kidney disease can also significantly reduce the survival probability of a COVID patient.

```
knitr::include_graphics("gfx/salinas_fig.png")
```

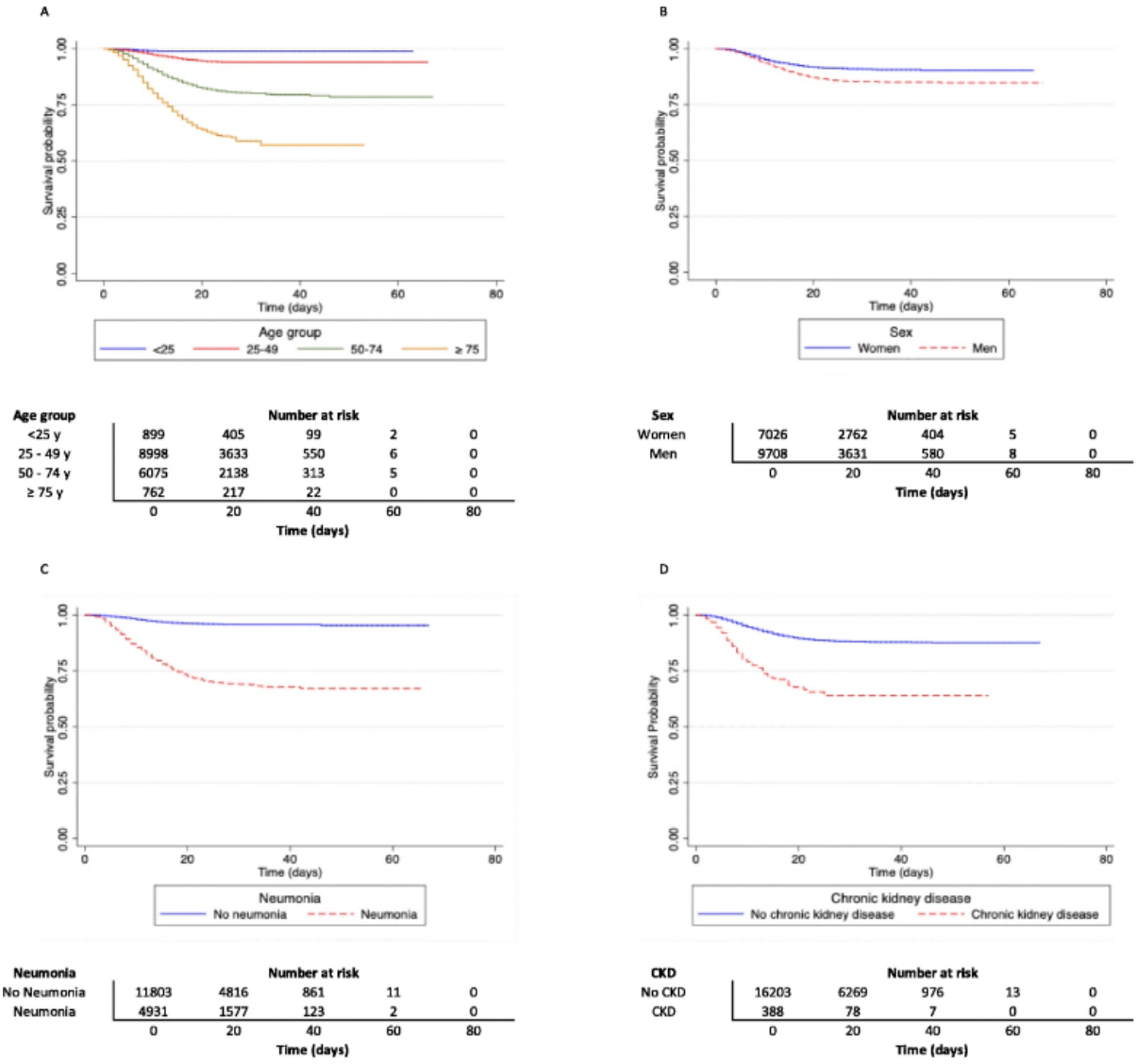


Figure 2: Displaying Kaplan-Meier curves from Figure 1 of Salinas-Escudero et. al. article

2.4 Cox Proportional Hazards Model

The Cox proportional hazards model is analogous to multiple linear regression. This model is a useful tool in survival analysis because it allows for the testing of the difference in survival times between particular groups while including other predictors (possible confounding variables) in the model [Bewick et al., 2004]. As we will see in the example in the application section, the R output for the proportional hazards model is very similar to standard linear regression.

The Cox model does not rely on any assumptions of the underlying distribution of the response variable. It does, however, assume that the hazard ratio does not depend on time. As noted by Spruance et al., "The hazard ratio is an estimate of the ratio of the hazard rate in the treated versus the control group... Thus, in a clinical trial where disease resolution is the endpoint, the hazard ratio indicates the relative likelihood of disease resolution in treated versus control subjects at any given point in time" [Spruance et al., 2004].

The formula for the Cox Proportional Hazards model with p covariates is given by,

$h(t, X) = h_0(t)\ exp(\sum_i^p \beta_i X_i)$,

where X_i are the covariates and β_i are the regression coefficients [Emmert-Streib and Dehmer, 2019].

2.4.1 Examples in Literature

Recall the study mentioned in section 2.3.1 done on COVID-19 patients in Mexico. Salinas-Escudero et al. created two separate Cox proportional hazards model for women and men. Below is a table with their results from the Cox model for Mexican women with COVID.

	Women			
	Hazard Ratio	***p*-value**	**95% CI for the Hazard Ratio**	
Age [a]				
25–49 y				
50–74 y	2.37	$p < 0.001$	1.88	2.99
75 + y	4.41	$p < 0.001$	3.32	5.87
Chronic kidney disease	1.90	$p < 0.001$	1.36	2.66
Pneumonia	2.09	$p < 0.001$	1.63	2.67
Intubation	2.83	$p < 0.001$	2.20	3.63
Hospitalization	6.57	$p < 0.001$	4.69	9.22
Health Services (Private, reference)				
IMSS Services	4.34	$p < 0.001$	2.28	8.26
ISSSTE Services	3.12	0.001	1.57	6.18
SS Services	3.22	$p < 0.001$	1.70	6.08
Other Public services	2.76	0.006	1.35	5.65

Figure 3: Displaying Cox Proportional Hazards Model from Table 2 of Salinas-Escudero et. al. article

```
knitr::include_graphics("gfx/salinas_table.png")
```

This table shows us all the significant covariates that their research found. The hazard ratios can be interpreted as follows:

- Women 75 and over had 4.41 times the risk of dying compared to women under 49,
- Women between the ages of 50-74 had 2.37 times the risk of dying compared to women under 49,
- Women with chronic kidney disease (CKD) had 1.9 times the risk of dying compared to women who do not have CKD,
- Women who needed to be hospitalized had 6.57 times the risk of dying compared to women who did not need to be hospitalized, etc [Salinas-Escudero et al., 2020].

3 Application to data

3.1 The dataset

Let's take a look at the `gbsg` data set from the survival package in R. It has data on 686 breast cancer patients from a trial conducted by the German Breast Cancer Study Group from 1984 to 1989 [Therneau, 2023]. The event in this case is a patient going out of remission *or* dying. and the time variable is time in days that a patient is in remission from breast cancer. Some of the other variables in this dataset we will be working with include:

- age: age of the patient in years
- meno: patient menopausal status (0 = premenopausal, 1 = postmenopausal)
- size: size of tumor in millimeters
- grade : grade of tumor on a scale between 1-3
- nodes: number of cancer positive lymph nodes
- pgr: progesterone receptors measured in fmol per liter
- hormon: patient hormonal therapy status (0 = did not receive hormonal therapy, 1 = received hormonal therapy)
- rfstime: recurrence free survival time; days to first of recurrence, death, or last follow up
- status: cancer status (0 = alive without recurrence, 1 = recurrence or death).

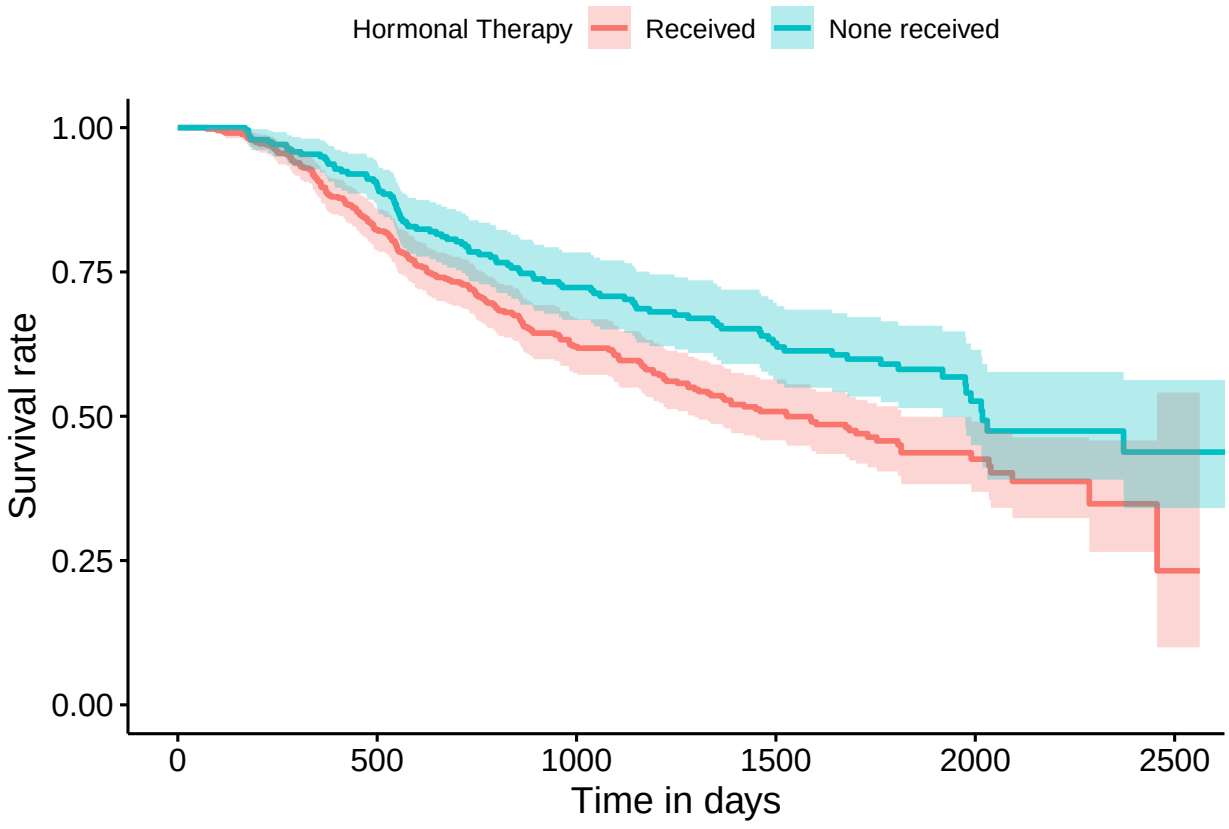
3.2 Kaplan-Meier Curves

Let's first take a look at the variable `hormon` which indicates if the patient was on hormonal therapy or not. We can plot the Kaplan-Meier curves for patients grouped by their therapy

status and compare their survival times. We can use the function ggsurvplot from the survminer package to help create plots for the KM curves.

```
# Generate fits for survival curves
km_htherapy <- survfit(Surv(rfstime, status) ~ hormon, data = gbsg)


# Plot both survival curves
ggsurvplot(km_htherapy, censor = FALSE, xlab = "Time in days",
           ylab = "Survival rate",legend.title = "Hormonal Therapy",
           legend.labs = c("Received", "None received"),
           conf.int = TRUE)
```



It appears that for patients who were on the hormonal therapy, the survival rate was higher for nearly all instances of time, t. But is this difference significant? We can use the

log rank test to evaluate the difference.

3.3 Log Rank Test

```
surv_diff <- survdiff(Surv(rfstime, status) ~ hormon, data = gbsg)

surv_diff
```

```
## Call:

## survdiff(formula = Surv(rfstime, status) ~ hormon, data = gbsg)

##

##             N Observed Expected (O-E)^2/E (O-E)^2/V

## hormon=0 440      205      180      3.37      8.56

## hormon=1 246       94      119      5.12      8.56

##

##  Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```

With a p-value of $p = 0.003 < 0.5$ we would reject the null hypothesis that the survival curves for the two groups are the same. There is enough evidence to suggest that the patients who were on hormonal therapy had better survival rates compared to those who did not.

3.4 Cox Proportional Hazards Model

Next we can build a model for the data using the Cox proportional hazards model. We can use this to explore more covariates than just hormone therapy.

```r
# changing from numeric to factor
breast_cancer <- gbsg %>%
  mutate(grade = as.factor(grade)
         , meno = as.factor(meno)
         , hormon = as.factor(hormon)) %>%
  select(-pid)

# cphm with all variables
mod1 <- coxph(Surv(rfstime, status) ~ ., data = breast_cancer)
summary(mod1)
```

```
## Call:
## coxph(formula = Surv(rfstime, status) ~ ., data = breast_cancer)
##
##   n= 686, number of events= 299
##
##                coef  exp(coef)   se(coef)      z Pr(>|z|)
## age      -0.0094592  0.9905854  0.0093006 -1.017 0.309126
## meno1     0.2584448  1.2949147  0.1834765  1.409 0.158954
## size      0.0077961  1.0078266  0.0039390  1.979 0.047794 *
## grade2    0.6361117  1.8891211  0.2492025  2.553 0.010693 *
## grade3    0.7796542  2.1807181  0.2684801  2.904 0.003685 **
## nodes     0.0487886  1.0499984  0.0074471  6.551  5.7e-11 ***
## pgr      -0.0022172  0.9977852  0.0005735 -3.866 0.000111 ***
## er        0.0001973  1.0001973  0.0004504  0.438 0.661307
```

```
## hormon1 -0.3462784  0.7073155  0.1290747 -2.683 0.007301 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age        0.9906     1.0095    0.9727    1.0088
## meno1      1.2949     0.7723    0.9038    1.8553
## size       1.0078     0.9922    1.0001    1.0156
## grade2     1.8891     0.5293    1.1591    3.0788
## grade3     2.1807     0.4586    1.2885    3.6909
## nodes      1.0500     0.9524    1.0348    1.0654
## pgr        0.9978     1.0022    0.9967    0.9989
## er         1.0002     0.9998    0.9993    1.0011
## hormon1    0.7073     1.4138    0.5492    0.9109
##
## Concordance= 0.692  (se = 0.015 )
## Likelihood ratio test= 104.8  on 9 df,   p=<2e-16
## Wald test            = 114.8  on 9 df,   p=<2e-16
## Score (logrank) test = 120.7  on 9 df,   p=<2e-16
```

Based on the results of the likelihood ratio test, Wald test, and log rank test, there is evidence to suggest the overall model is statistically significant. It appears that the variables that have a significant effect on survival time are `size`, `grade`, `nodes`, `pgr`, and `hormon`. We can create a new model using only these significant covariates. The `exp(coef)` column gives us the hazard ratio. We can use the fit2df() function. Table 1 below gives us

the hazard ratio, 95% confidence interval and p-value for each variable.

```
# cphm with selected variables
mod2 <- coxph(Surv(rfstime, status) ~ size + grade + nodes + pgr + hormon,
              data = breast_cancer)


#table
mod2 %>%
  fit2df(condense = FALSE) %>%
  kable(digits = 4)
```

explanatory	HR	L95	U95	p
size	1.0073	0.9997	1.0150	0.0601
grade2	1.9041	1.1688	3.1022	0.0097
grade3	2.1995	1.3001	3.7211	0.0033
nodes	1.0502	1.0350	1.0657	0.0000
pgr	0.9978	0.9967	0.9989	0.0001
hormon1	0.7236	0.5655	0.9260	0.0101

From Table 1 we can see that in this model size is no longer significant. The 95% CI contains 1 and the p-value is $p = 0.0601 > 0.05$. As for the effect of the grade of the cancerous tumor, it appears that having a tumor of grade 2 or higher is strongly associated with an increased risk of cancer recurrence or death. A greater number of cancer positive lymph nodes also appears to be associated with higher risk. Conversely, it appears that higher levels of progesterone receptor proteins is associated with a decreased risk. Patients who received hormone therapy appeared to have lower risk of cancer recurrence or death compared to patients who did not receive hormone therapy.

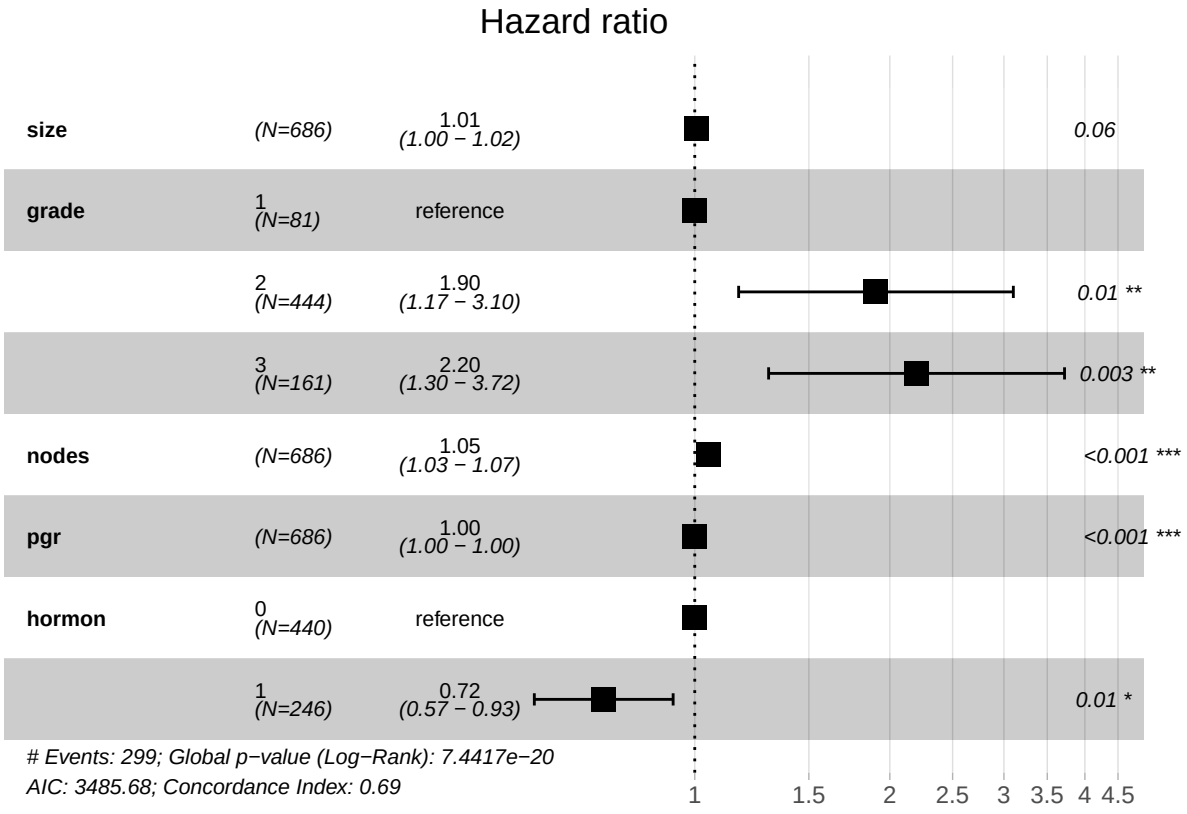
We can also use the ggforest() function in the survminer package to visually represent the hazard ratio and confidence intervals [Kassambara et al., 2021].

```
mod2 %>%

  ggforest()
```

```
## Warning in .get_data(model, data = data): The ‘data‘ argument is not provided.

## Data will be extracted from model fit.
```



This visualization is helpful in conveying how great an effect grade of tumor and hormone therapy have on the risk of breast cancer recurrence or death.

References

Viv Bewick, Liz Cheek, and Jonathan Ball. *Critical Care*, 8(5):389, 2004. ISSN 1364-8535. doi: 10.1186/cc2955. URL http://dx.doi.org/10.1186/cc2955.

T G Clark, M J Bradburn, S B Love, and D G Altman. Survival analysis part i: Basic concepts and first analyses. *British Journal of Cancer*, 89(2):232–238, July 2003. ISSN 1532-1827. doi: 10.1038/sj.bjc.6601118. URL http://dx.doi.org/10.1038/sj.bjc.6601118.

Frank Emmert-Streib and Matthias Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038, September 2019. ISSN 2504-4990. doi: 10.3390/make1030058. URL http://dx.doi.org/10.3390/make1030058.

Alboukadel Kassambara, Marcin Kosinski, and Przemyslaw Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*, 2021. URL https://CRAN.R-project.org/package=survminer. R package version 0.4.9.

D. G. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Springer, 2020.

Guillermo Salinas-Escudero, María Fernanda Carrillo-Vega, Víctor Granados-García, Silvia Martínez-Valverde, Filiberto Toledano-Toledano, and Juan Garduño-Espinosa. A survival analysis of covid-19 in the mexican population. *BMC Public Health*, 20 (1), October 2020. ISSN 1471-2458. doi: 10.1186/s12889-020-09721-2. URL http://dx.doi.org/10.1186/s12889-020-09721-2.

Spotswood L. Spruance, Julia E. Reid, Michael Grace, and Matthew Samore. Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, 48(8):2787–2792, August 2004.

ISSN 1098-6596. doi: 10.1128/aac.48.8.2787-2792.2004. URL http://dx.doi.org/10.1128/AAC.48.8.2787-2792.2004.

Terry M Therneau. *A Package for Survival Analysis in R*, 2023. URL https://CRAN.R-project.org/package=survival. R package version 3.5-7.