STAD92H3 Assessement, Supervised by Dr. Shahriar Shams

Survival Analysis on Cancer Data

Dewei Lin

September 1, 2023

Abstract

This project applies Survival Analysis techniques to survival data. The report covers EDA (explanatory data analysis), answers to instructor's questions, an R code appendix, and references.

1 EDA

1.1 Data Description

The dataset, consisting of 914 entries, provided contains the following information about some cancer patients:

- **id**: Patient ID
- **f_time**: Follow-up time
- **vstatus**: Dead (event), Alive (Censored)
- **t_stage**: Tumour stage (I to V)
- **ulcer**: Presence of ulcer (0 = No, 1 = Yes)
- **thick**: Thickness of tumour (in mm)

1.2 Missing Value

The dataset contains many missing values that needs to be addressed, which is presented in the following table.

Variable	Missing Proportion	Missing Cases
id	0.0%	0
f_time	0.0%	0
vstatus	0.0%	0
t_stage	2.2%	20
ulcer	25.8%	236
thick	0.0%	0

A small portion of **t_stage** values are missing. Due to our limited understanding of the missing data mechanism and the low count of missing values, a possible approach for subsequent analysis could involve excluding records with missing **t_stage** values.

Due to the substantial proportion of missing values in the **ulcer** variable and the lack of knowledge about the underlying missing data mechanism, I will omit this variable when fitting the survival model.

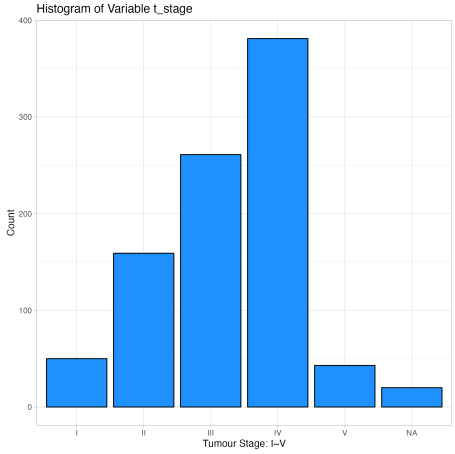


Figure 1: Histogram of Variable **t_stage**

1.3 Potential Errors

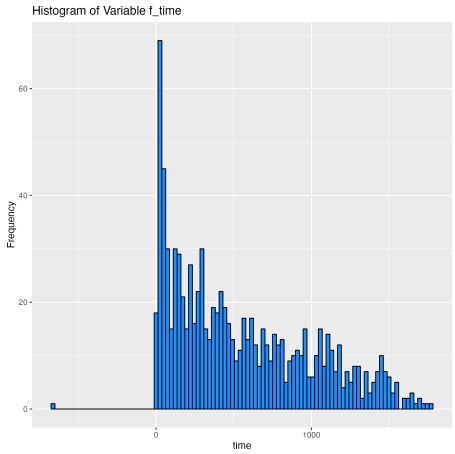


Figure 2: Histogram of **f_time**

Negative survival time **f_time** values have been identified. To rectify this error, I will remove them during subsequent analyses.

2 KM Survival Curves and the Log-Rank Test

To better understand the survival mechanism and answer **(Question 1a)**, the plot of KM survival curve of the entire data, along with 95% CI are shown as follows: The median survival time

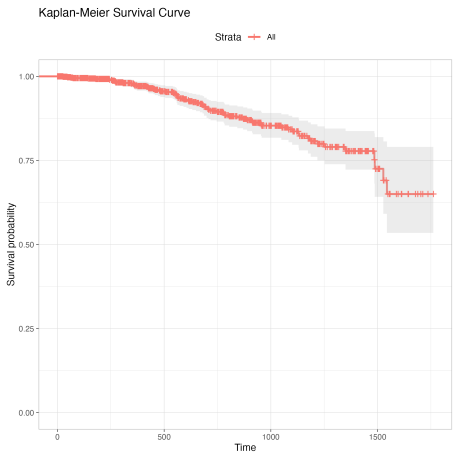


Figure 3: KM Survival Curve for entire data

in this dataset is not reached, hence we are not able to calculate median survival time and the corresponding 95% CI.

```
Call: survfit(formula = Surv(f_time,
event)~1, data = sur.data)

      n events median 0.95LCL 0.95UCL
[1,] 914     73     NA      NA      NA
```

We are interested in investigating the difference between patients either with presence of **ulcer** or not in terms of survival. **(Question 1b)** A plot that includes survival curves for both groups is in Figure 4:

Note that the yellow curve stands for no **ulcer** group while blue curve stands for patients with presence of **ulcer**. Graphically speaking, there is significant difference between two groups in terms of survival. However, I will conduct a log-rank test to answer the question:

```
Call:
survdiff(formula = Surv(f_time, vstatus)
~ ulcer, data = work_d)

          N   O    E  (O-E)^2/E (O-E)^2/V
ulcer=0 452  15 33.8   10.4      33.9
ulcer=1 226  34 15.2   23.2      33.9

Chisq=33.9 on 1 degrees of freedom,p=6e-09
```

The p-value associated with log-rank test is extremely small, we reject the H_0 and conclude

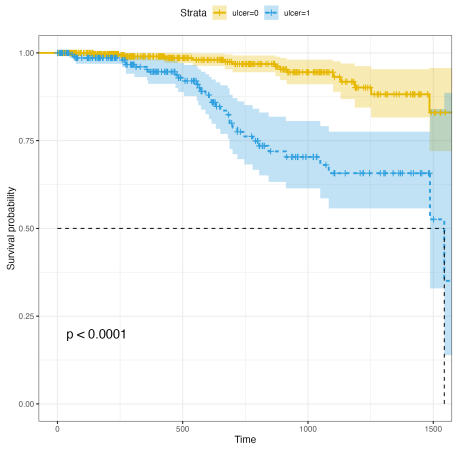


Figure 4: KM Survival Curve for 2 Groups ((0=No, 1=Yes))

that there is significant difference between two **ulcer** groups in terms of survival.

3 Data Pre-Processing

Before fitting survival models, I'd like to clarify how I pre-processed data.

Number of Cases	Reason
1	`f_time` < 0
20	Missing

Table 1: Removed cases

4 The Semi-Parametric Cox PH Model

The basic formula for the Cox Proportional Hazard model can be written as:

$$h(t|x) = h_0 \exp\{x'\beta\}$$

In this context:

- $h(t|x)$ is the predicted value of the hazard rate for a specific x.
- h_0 is an arbitrary and unspecified baseline hazard function.
- $x'\beta$ represents all specified risk factors or variables multiplied by their corresponding coefficient values.

Among the three available variables, I utilized the Cox Proportional Hazards model and integrated the predictors **t_stage** and **thick**. Due

to a substantial amount of missing data, the variable **ulcer** was omitted from the analysis. Consequently, the formulation of the Cox PH model is as follows **(Question 2a)** :

$$h(t|x) = h_0 \exp\left\{\begin{array}{l} 0.09631 \cdot \text{thick} \\ -0.62746 \cdot \text{t\_stage\_II} \\ +0.50731 \cdot \text{t\_stage\_III} \\ +1.39539 \cdot \text{t\_stage\_IV} \\ +1.75615 \cdot \text{t\_stage\_V} \end{array}\right\}$$

Model Interpretations

The interpretation of the model:

For **thick**, $\exp(0.09631) = 1.10110$. This indicates an approximately 10.11% increase in the expected hazard relative to a 1mm increase in the thickness of the tumor, while holding everything else constant.

For **t_stage**, I will interpret the coefficient in terms of a one-level increase from the lower level:

- If the tumor stage of the patient changes from I to II, then we expect an approximately 46.61% decrease in hazard in terms of the occurrence of death.
- If the tumor stage of the patient changes from II to III, then we expect an approximately 66.08% increase in hazard in terms of the occurrence of death.
- If the tumor stage of the patient changes from III to IV, then we expect an approximately 303.7% increase in hazard in terms of the occurrence of death.
- If the tumor stage of the patient changes from III to IV, then we expect an approximately 479% decrease in hazard in terms of the occurrence of death.

Model Diagnosis

The Cox PH model is valid under PH assumptions. The mathematical formulation of PH assumption can be expressed as:

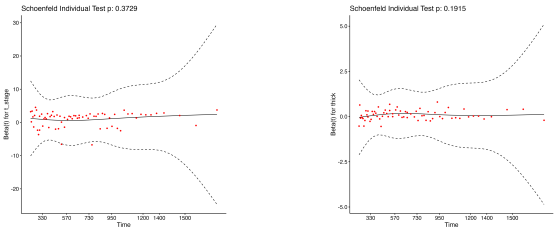
$$\hat{HR} = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} \tag{1}$$

$$= \frac{\hat{h_0} \exp\left[\sum \beta_i \hat{X}_i^*\right]}{\hat{h_0} \exp\left[\sum \beta_i \hat{X}_i\right]} \tag{2}$$

$$= \exp\left[\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\right] \tag{3}$$

The PH assumption holds if the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time. Simply put, the assumption holds when $\hat{HR}$ is constant over time

PH assumption is an be checked using statistical tests and graphical diagnostics based on the scaled Schoenfeld residuals. Under H_0, the PH assumption holds **(Question 2b)**.



(a) Schoenfeld Test for Variable **t_stage**

(b) Schoenfeld Test for Variable **thick**

Figure 5: Overall Caption for the Two Graphs

Graphically speaking, most of scaled Schoenfeld residuals falls into 95% CI and we can do statistical test to confirm our findings:

	chisq	df	p
thick	1.71	1	0.19
t_stage	4.25	4	0.37
GLOBAL	6.66	5	0.25

The p-value for predictors and global model is very large, we fail to reject the null hypothesis and conclude that the PH assumption holds. In conclusion, we are safe to apply Cox PH model.

5 The Parametric Model

All predictors included

For this part **(Question 3a)** , I included all three predictors in parametric models with Exponential, Wei-bull, Gamma, Log-normal, Log-logistic distribution as support of underlying

hazard function. As a result, the log-logistic distribution fits the data best based on AIC criteria (Table 2).

Distribution	AIC	Assumption
Log-logistic	908.1866	PO
Gamma	909.0833	AFT
Weibull	909.4518	AFT
Log-normal	912.1468	AFT
Exponential	930.3147	PH

Table 2: AIC values for different distributions

The Log-logistic hazard:

$$h(t) = \frac{\lambda p t^{p-1}}{1 + \lambda t^p}$$

where p is the shape parameter, λ is the scale parameter.

Best model

As mentioned earlier, we would prefer to remove **ulcer** from our model since there are too many missing values for this variable. **(Question 3b)**As a result, the best fit distribution becomes Wei-bull (Table 3).

Distribution	AIC	Assumption
Weibull	1329.846	AFT
Log-logistic	1331.065	PO
Gamma	1331.630	AFT
Log-normal	1341.730	AFT
Exponential	1358.884	PH

Table 3: AIC values for different distributions

Wei-bull AFT model can be expressed as:

$$S(t) = \exp\{-\lambda t^p\}, \hat{p} = 1.8, \quad \hat{\lambda} = 5490 \tag{4}$$

$$t = [-\ln S(t)]^{1/1.8} \exp \left\{ \begin{array}{l} -0.0548 \cdot \text{thick} \\ +0.349 \cdot \text{t\_stage\_II} \\ -0.291 \cdot \text{t\_stage\_III} \\ -0.764 \cdot \text{t\_stage\_IV} \\ -0.944 \cdot \text{t\_stage\_V} \end{array} \right\} \tag{5}$$

Model interpretations

For **thick**, the accelerating factor $\gamma_i = \exp\{-0.0548\} = 0.948$. This means that the median survival time is decreased by a factor of 0.948 for every one millimeter increase in the thickness of the tumor.

For **t_stage**, I will interpret the coefficient in terms of a one-level increase from the lower level:

- If the tumor stage of patients changes from I to II, we expect the median survival time to increase by a factor of 1.42.
- If the tumor stage of patients changes from I to II, we expect the median survival time to decrease by a factor of 0.748.
- If the tumor stage of patients changes from II to III, we expect the median survival time to decrease by a factor of 0.466.
- If the tumor stage of patients changes from III to IV, we expect the median survival time to decrease by a factor of 0.389.

6 Brief Summary

There are some findings based on previous three questions **(Question 4)**.

First, the median survival time is undefined for this dataset because the time to death for cancer can last much longer than a observational study can take.

Second, the survival can differ greatly given the presence of ulcer or not. However, there are a large proportion of missing value for this confounding variable and we can't make a assured conclusion.

Third, the odds of survival decreases with thickness of tumour increases. At a baseline of tumour stage II, higher levels of tumour stages are associated with higher risk in terms of death.

Acknowledgements

I would like to extend my sincere appreciation to my supervisor, Dr. Shams, for generously dedicating time from their vacation to impart valuable knowledge on survival analysis. I am also thankful to my fellow peers who have contributed to this reading course, making the learning experience collaborative and enriching.

References

[1] Alexandra Sciocchetti, Survival Analysis of Heart Failure Patients: Student Report, `https://rpubs.com/asciocchetti/survivalanalysis_heartfailure`.

[2] Kleinbaum, David G., and Mitchel Klein. "Survival Analysis: A Self-Learning Text." Third Edition. Springer, 2012.

[3] Moore, Dirk F. "Applied Survival Analysis Using R."

[4] Web sources on how to implement flexsurvreg()

Appendix: R codes

```
### Import data and libraries
```{r}
library(readxl)
library(survival)
library(survminer)
library(epiR)
library(purrr)
library(dplyr)
library(flexsurv)
library(ggplot2)
library(tidyr)
library(mice)
library(scales)
sur.data <- read_excel("~/Desktop/STAD92
/D92 Final Project
/STAD92_S2023_Project_Data.xlsx")
sur.data <- sur.data %>%
  mutate(event = ifelse(vstatus == "Dead",
  1, 0))
  %>% mutate(t_stage = factor(t_stage))
head(sur.data)
```

EDA

Data size

```{r}
nrow(sur.data)
```

Missing Value

```{r}
missing_prop <- data.frame(
Missing_Proportion=
percent(
colSums(is.na(sur.data))/nrow(sur.data)),
Missing_Cases=colSums(is.na(sur.data)))
missing_prop
```

Variable t_stage

There is a small proportion of values
of t_stage being missing. Because we
are not aware of the missing
mechanism and there are only few
missingvalues, we can choose to
remove data with missing
t_stage in future analysis.

```{r}
```
```

```
plot1 <- ggplot(sur.data,
aes(x = t_stage)) +
  geom_bar(fill = "dodgerblue",
  color = "black") +
  xlab("Tumour Stage: I~V") +
  ylab("Count") +
  ggtitle("Histogram of Variable t_stage")) +
  theme_light()
#ggsave("Histogram of Variable t_stage.png",
plot = plot1)
```

Variable ulcer

Because a large proportion of values of variable ulcer are missing and we are not aware of some missing mechanism, we have to exclude this variable when fitting the survival model.

Address Potential Errors

```{r}
plot2 <- ggplot(sur.data, aes(x = f_time)) +
  geom_histogram(fill = "dodgerblue",
  color = "black", bins = 100) +
  xlab("time") +
  ylab("Frequency") +
  ggtitle("Histogram of Variable f_time")
#ggsave("Histogram of Variable f_time.png",
plot = plot2)
```

```{r}
sum(sur.data$f_time<0)
sum(sur.data$f_time<0)/nrow(sur.data)
```

There exists negative f_time values which are not supposed to exist. I will remove the error in future analysis.

Q1

a)

Draw a KM survival curve for the entire data(without considering any independent variable) . Also calculate median survival time along with 95% confidence interval.

KM Survival Curve

```{r}
km_fit <- survfit(Surv(f_time, event) ~ 1,
data = sur.data)
KMSC <- ggsurvplot(
  km_fit,
  data = sur.data,
  conf.int = TRUE,
  ggtheme = theme_light(),
  xlim = c(0,max(sur.data$f_time)),
  surv.median.line = "hv",
  title = "Kaplan-Meier Survival Curve"
) +
plot3 <- KMSC$plot
#ggsave("KM Survival Curve.png",
plot = plot3)
plot3
```

```{r}
sum(sur.data$event)/nrow(sur.data)
```

Median Survival Time, and 95% CI

```{r}
km_fit
```

b)

Draw another KM curve that shows two survival curves (two curves in one plot): one curve for ulcer = No, and another curve for ulcer=Yes. Also carry out a log-rank test to compare the survival between these two categories of ulcer presence. Interpret your outputs.

KM Survival Curve for 2 Groups

```{r}
work_d <- na.omit(data.frame(
sur.data["f_time"],
sur.data["event"],
sur.data["ulcer"]))
km_fit <- survfit(Surv(f_time, event)
~ ulcer, work_d)
plot.b <- ggsurvplot(km_fit,
    pval = TRUE, conf.int = TRUE,
    risk.table.col = "strata",
    linetype = "strata",
    surv.median.line = "hv",
    ggtheme = theme_bw(),
    palette = c("#E7B800",
    "#2E9FDF"),
    work_d)
plot4<-plot.b$plot
#ggsave("KM Survival Curve for
```

```
2 Groups.png",
plot = plot4)
plot4
```

Log-Rank Test

Conduct Log-Rank Test:

H_0: No difference between two ulcer groups in terms of survival

```{r}
log.rank <- survdiff(Surv(f_time, event)
~ ulcer, work_d)
log.rank
```

Comment:\
The p-value associated with log-rank test is greater than 0.05, we fail to reject the H_0 and conclude that there is no significant difference between two ucler groups in terms of survival.

\newpage

Preprocess data

```{r}
work_d <- sur.data %>%
  filter(f_time > 0, !is.na(t_stage))
```

Q2

a)

Fit a Cox-PH model using the given data. Use t_stge, ulcer and thick as independent variables. You can remove independent variables from the model if you find necessary. Interpret the outputs of your model.

Fit a Cox PH model

```{r}
cox.ph <- coxph(Surv(f_time, event) ~
thick+t_stage, work_d)
summary(cox.ph)
```

Interpret the model

TBD

b)

Assess the proportional hazard assumption for each of the independent variables. What conclusion do you make in terms of the applicability of the Cox-PH model.

Statistical Test: cox.zph

The cox.zph function will test proportionality of all the predictors in the model by creating interactions with time using the transformation of time specified in the transform option.

H_0: Time-Invariant

```{r}
test.ph <- cox.zph(cox.ph)
test.ph
```

Graphical Test

```{r}
cox.plot <- ggcoxzph(test.ph)
plot5 <- cox.plot$'1'
plot6 <- cox.plot$'2'
plot5;plot6
#ggsave("Schoenfeld Test
for Variable thick.png", plot = plot5)
#ggsave("Schoenfeld Test for
Variable t_stage.png", plot = plot6)
```

\newpage

Q3

a)

```{r}
par_fits <- tibble(
  dist_param = c("exp", "weibull",
  "gamma", "lognormal", "llogis"),
  dist_name = c("Exponential",
  "Weibull", "Gamma",
"Log-normal", "Log-logistic")
) %>%
  mutate(
    fit = map(dist_param,
    ~flexsurvreg(Surv(f_time, event)
    ~ thick+t_stage, data = work_d,
    dist = .x)),


```

```
    fit_smry = map(fit, ~summary(.x,
    type = "hazard", ci = FALSE,
    tidy = TRUE)),
    AIC = map_dbl(fit, ~.x$AIC)
  )
par_fits %>% arrange(par_fits$AIC)
%>% select(dist_name, AIC)
```

```{r}
flexsurvreg(Surv(f_time, event) ~
thick +t_stage, data = work_d,
dist = "weibull")
```