Modeling of Cancer Patient: A Survival Analysis Approach

CS 510 PROJECT WORK

---

by

Isaac Nwi - Mozu

Fall 2020

---

1 Introduction

Survival analysis is a collection of statistical methods that are used to describe, explain, or predict the occurrence and timing of events. The basic functions of survival analysis are the same in all fields, but the name usually depends on the field of study. For instance in engineering it is called reliability analysis, in sociology it is known as event history analysis, in economics it is famous with the name of duration analysis and medical researchers give it the name of survival analysis.

In the classical mathematical modeling, data are collected over a finite period of time. For an example, modeling population increase from census data, predicting interest gain and so forth. However, in some cases certain features of the individual under study may not be observed for all the individuals in our study population. This creates an irregularity/noise in our data set. The second special feature is that the data always follows a skewed distribution.

In many dynamical systems, it is common that the amount of follow-up for the individuals in a sample vary from subject to subject. We may have study dropout. The combination of staggered observation into the study and differential follow-up creates some unusual difficulties in the analysis of such data that cannot be handled properly by the standard mathematical/statistical methods such as ordinary regression model. For illustration of this phenomenon, consider the figure below that study cancer pa-

tients for eight weeks until patient's die with the cancer. Here, our main focuses are death of patients. The patients may have staggered entries into the study i.e., the patient's have different entry time. Patients A, D and E started from the beginning of the study but the investigator was not able to track patient E and so was lost to follow up before the end of the whole study. Patient C started in the middle of the study and later did not show interest in the study and so dropped out.

In this study, we show some systematic approaches that can be used to model such problems. We also hope that by the end of this study, clinicians will be able to find an efficient way to model survival data.

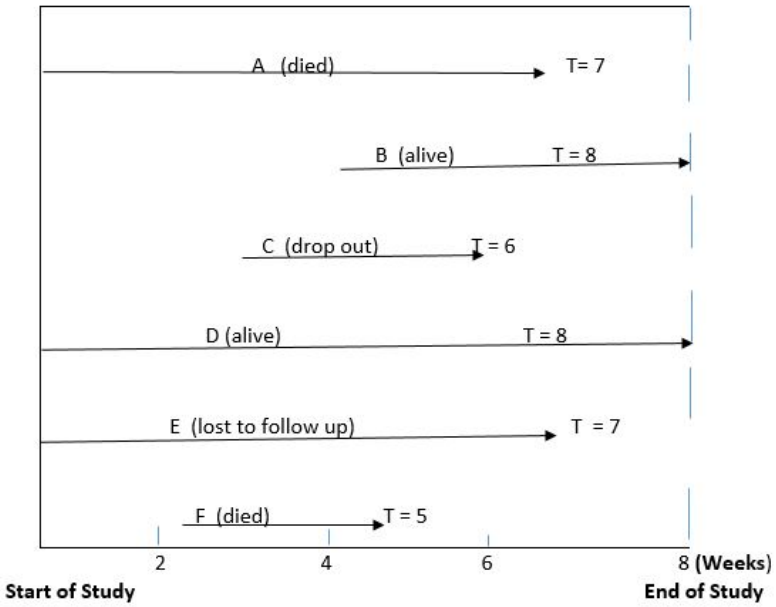

Figure 1: Illustration of types of censoring

1.1 Materials and method

The data contains 199 cancer patients who reported to a hospital at different date (irregularity in the time) who were diagnosed within 9 year period. The following samples (predictors) were recorded from the patients is the stage of cancer. Patients were followed for one year (365 days) and the number of days until a patient died was recorded

Various form of distribution were tested to model the survival rate and the death rate of the patient. Our purpose is to predict how long a patient survive or die with the disease.

The following censoring assumptions on patient were made.

• Patients who were alive disease-free were considered censored.

• Patients who died from other causes than cervical cancer were considered censored.

• Patients who were still alive with disease as at their last follow up date were considered to be censored.

• Patients who were lost to follow-up were considered to have been censored.

2 Methodology and mathematical formulation

The definition of the distributions considered in this study was taken from Tableman and Kim (2003) as follows:)

Let T be the random variable representing survival time of a patient, the survival function $S(t)$, is defined as the probability that an individual survive beyond time t. It is given by:

$$S(t) = P(T > t), 0 \leq S(t) \leq 1 \tag{1}$$

2.1 Properties of survival function

1. It is decreasing function.

2. At time t = 0, S(t)=1

3. At time $t = \infty, S(t) = 0$. As time goes to infinity, the survival curve goes to 0.

If the failure time is continuous random variable then:

$$S(t) = \int_t^{\infty} f(u)du, \forall \text{t} \in [0, \infty] \tag{2}$$

2.1.1 Relationship between survival function and cumulative distribution function

Let $f(t), t \geq 0$, denote the probability density function (pdf) of T and $F(t) = P(T \leq t) = \int_0^t f(u)du, t \geq 0$ be the cumulative distribution function (cdf) of T, then

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= 1 - F(t) \end{aligned} \tag{3}$$

Taking the differential of both sides

$$\begin{aligned} S'(t) &= -F(t) \\ &= -f(t) \end{aligned} \tag{4}$$

2.1.2 The hazard function

The hazard function λ (t), is defined mathematically as:

$$\lambda(t) = \lim_{\nabla t \to 0} \frac{P(\text{an individual who survive to time t fail in}(t, t + \nabla t))}{\nabla t} \tag{5}$$

λ(t) has two major properties: it is always nonnegative and has no upper bound. The cumulative hazard Λ, and the hazard function $\lambda(t)$, for continuous random variable are related by:

$$\Lambda = \int_0^t \lambda(u) du \tag{6}$$

2.1.3 Continuous random variable of survival function

If the failure time is continuous then:

$$\begin{aligned}\lambda(t) &= \lim_{\nabla t \to 0} \frac{1}{\nabla} P(t \leq T < t + \nabla t | T \geq t) \\ &= \lim_{\nabla t \to 0} \frac{1}{\nabla t} \frac{P([t \leq T < t + \nabla t] \cap [T \geq t])}{P(T \geq t)} \\ &= \lim_{\nabla t \to 0} \frac{1}{\nabla t} \frac{P([t \leq T < t + \nabla t])}{P(T \geq t)} \\ &= \frac{f(t)}{S(t)}\end{aligned} \tag{7}$$

From equation (5)

$$\lambda(t) = \frac{-S'(t)}{S(t)} \tag{8}$$

$$= \frac{-d(logS(t))}{dt} \tag{9}$$

Integrating both sides:

$$\Lambda(t) = -logS(t) \tag{10}$$

$$S(t) = \exp^{-\Lambda(t)} \tag{11}$$

$$S(t) = \exp^{-\int_0^t \lambda(t)dt} \tag{12}$$

Equation (14) gives the relationship between cumulative hazard and the survival function whiles equation (15) gives the relationship between hazard function and the survival function. Thus given any one of three functions S(t), $\Lambda(t)$ and f(t) , the others two can be derived.

2.2 The Exponential Distribution

The exponential distribution is a parametric distribution with only one parameter. In exponential distribution, hazard function is assumed to be constant over time.

Consider the density function f(t), given by:

$$f(t) = \lambda \exp(-\lambda t) \tag{13}$$

Then for eqn(3.46) above we have the survival function as:

$$\begin{aligned} S(t) &= \int_0^\infty f u du \\ &= \exp^{-\lambda t} \end{aligned} \tag{14}$$

Taking log of both side of eqn(2):

$$\log S(t) = -\lambda t \tag{15}$$

A graph of survival estimate of $\log S(t)$ against time t, should produce approximately straight line with negative slope if the exponential assumption or model is valid.

The associated hazard function is given as:

$$\begin{aligned} \lambda(t) &= \frac{f(u)}{S(t)} \\ &= \lambda \end{aligned} \tag{16}$$

From eqn(3.49), the cumulative hazard is estimated as:

$$\begin{aligned} \Lambda(t) &= -\log S(t) \\ &= \lambda t \end{aligned} \tag{17}$$

2.3 The Weibull Distribution

The Weibull model is a parametric distribution with two parameters. The Weibull model is a slight modification of the exponential model. The model is obtain if the error term ϵ has an extreme value distribution with the following density: $f_\epsilon(x) = \exp(x - \exp^x), -\infty < x < \infty$

Equivalently, T has the Weibull distribution with the following density:

$f(t) = \lambda k(\lambda t)^(k-1)\exp(-(\lambda t)^k), t \geq 0$

Then from from eqn (3.5) we have the survival function as:

$$\begin{aligned} S(t) &= \int_0^\infty f(u)du \\ &= \exp(-(\lambda t)^k) \end{aligned} \tag{18}$$

The above eqn (3.55) gives the survival function for a survival data that follows a Weibull model. It is also used to check if the Weibull model is appropriate as follows. Taking log of both side of eqn (3.55):

$$\log S(t) = -(\lambda t)^k \tag{19}$$

$$\log(-\log S(t)) = k\log\lambda + k\log \tag{20}$$

A straight line in the plot of $\log(-\log S(t))$ vs. $\log t$ indicates a Weibull model. An (approximate) straight line indicates the Weibull model is a reasonable choice for

the data.

The hazard function is given as:

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} \\ &= \lambda k(\lambda t)^{k-1} \end{aligned} \tag{21}$$

From eqn (3.58) above, the cumulative hazard is estimated as:

$$\begin{aligned} \Lambda(t) &= -\log S(t) \\ &= (\lambda t)^k \end{aligned} \tag{22}$$

λ - the scale parameter

k - the shape parameter.

The Weibull distribution is convenient because of its simple form. it includes several hazard shapes:

k = 1 means constant hazard

$0 < k < 1$ implies decreasing hazard

$k > 1$ implies increasing hazard

The Weibull distribution reduces to exponential distribution when k = 1.

The λ is a scale parameter in that the effect of different values of λ is just to change the scale on the horizontal (t) axis, not the basic shape of the graph. This model is very flexible and has been found to provide a good description of many types of time-to-event data.

We might expect an increasing Weibull hazard to be useful for modeling survival

times of leukemia patients not responding to treatment, where the event of interest is death. As survival time increases for such a patient, and as the prognosis accordingly worsens, the patient's potential for dying of the disease also increases.

We might expect some decreasing Weibull hazard to well model the death times of patients recovering from surgery.

2.4 Log-Logistic Distribution

This model has become popular, for like the Weibull, it has simple algebraic expressions for the survivor and hazard functions. Hence, handling censored data is easier than with the log-normal while providing a good approximation to it except in the extreme tails

The lifetime T is log-logistically distributed if

$Y = \log(T)$ is logistically distributed with location parameter μ and scale parameter γ. Hence, Y is also of the form

$Y = \mu + \gamma Z$ where Z is a standard logistic random variable with density

$$\frac{\exp(z)}{(1+\exp(z))^2}, -\infty < z < \infty$$

Equivalently, T has the log-logistic distribution with the following density:

$$f(t) = \lambda k(\lambda t)^{k-1}(1+(\lambda t)^k)^{-2}, t \geq 0, k > 0, \lambda > 0 \tag{23}$$

We have the survival function as

$$\begin{aligned} S(t) &= \int_0^\infty f(u)du \\ &= \frac{1}{1+(\lambda t)^k} \end{aligned} \tag{24}$$

Equation (3.64) can further be simplified as:

$$\frac{S(t)}{1-S(t)} = (\lambda t)^{-k} \tag{25}$$

Taking log of both side of eqn (3.66) gives:

$$-\log(\frac{S(t)}{1-S(t)}) = k\log\lambda + k\log t \tag{26}$$

Similarly, A straight (approximately straight line) line in the plot of $\log t$ against $-\log(\frac{S(t)}{1-S(t)})$ indicates a log-logistics model. We can use the above equation to check if the log-logistics model is a reasonable choice for the survival time given a data set.

The hazard function is given as:

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\lambda k(\lambda t)^{k-1}}{1+(\lambda t)^k} \end{aligned} \tag{27}$$

From eqn (10) above, the cumulative hazard is estimated as:

$$\begin{aligned} \lambda &= -\log S(t) \\ &= -\log[\frac{1}{1+(\lambda t)^k}] \end{aligned} \tag{28}$$

2.5 Model Fit Selection

One can fit various survival functions to the data and visually compare how similar the survival functions are to the a non-parametric model estimate of the survival function. In that case, the graph of the distribution that is nearest to that of the non-parametric estimate is the best model.

Another is to use the Akaike Information Criterion (AIC) to choose the closest model. $AIC = -2Log(maximumlikelihood) + k \times p$, whrer p is number of parameters in each model under consideration and k a predetermined constant. The lower the AIC the better the model, Diez (2012).

3 Results

4 Parametric Model:

Table 1 below summarized the AIC for each parametric model commonly use. The AIC result suggested that the Weibull distribution is the best for estimating the survival time for fitting the data.

Table 1: Model selection using AIC

	Exponential	Log - Logistics	Weibull
AIC values	561.2192	541.4968	541.4593

A graph of log(t) against log(-log(S(t)) should produce an approximate straight line if the Weibull assumption is true. Where t is the survival time and S(t) is the survival of time. Figure 2 below produces an approximate straight line. This shows that the Weibull distribution is appropriate.

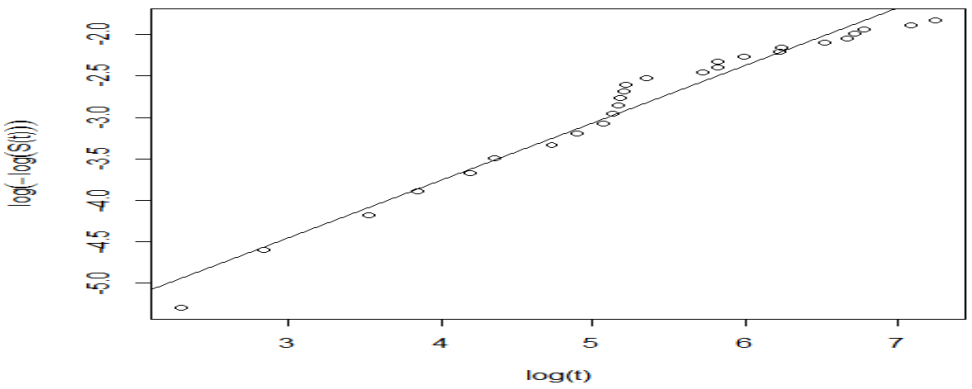

Figure 2: Weibull assumption graph for no predictor

Figure 3 below shows the survival graph for the Weibull. The survival curve decreases downwards with time. Patient tend to have higher survival as the least survival estimated by this graph is 75%.

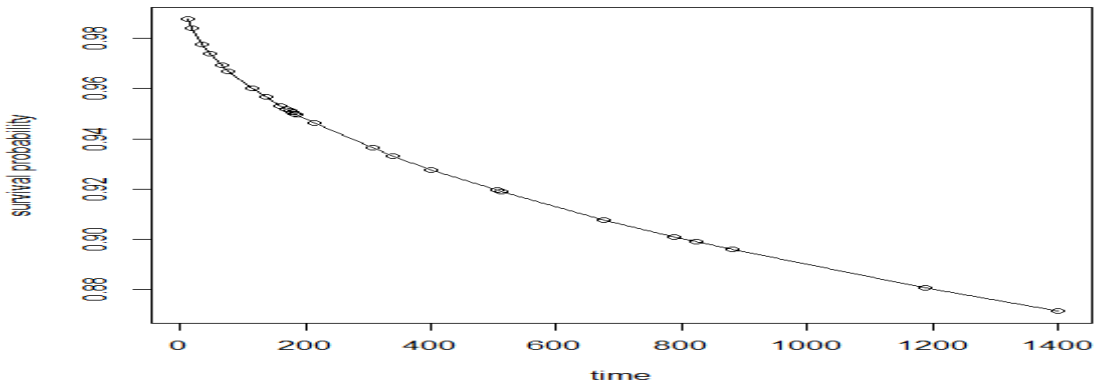

Figure 3: Weibull survival graph for no predictor

Figure 4 shows the Weibull hazard. Again, it was observed that the hazard increases slowly with time.

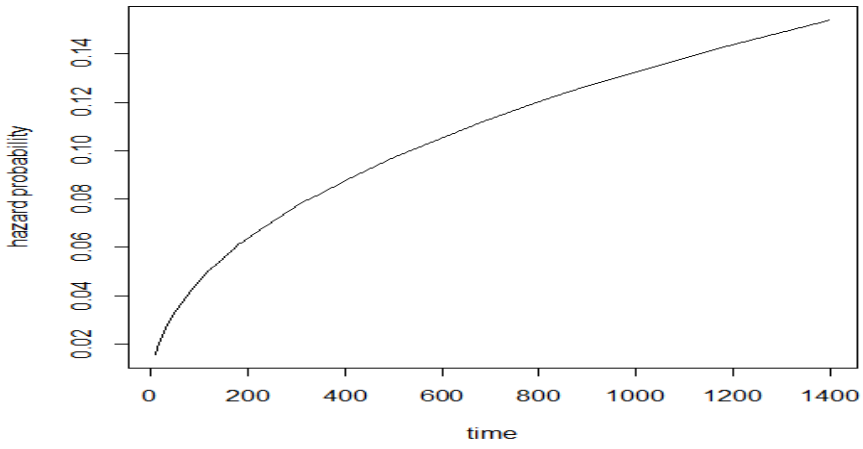

Figure 4: Weibull hazard graph for no predictor

The Weibull model was validated by plotting both the Weibull and the non-parametric model survival function on the same graph. Again if the Weibull model is perfect, then two graphs must go together. From figure 5, the two graphs move together.

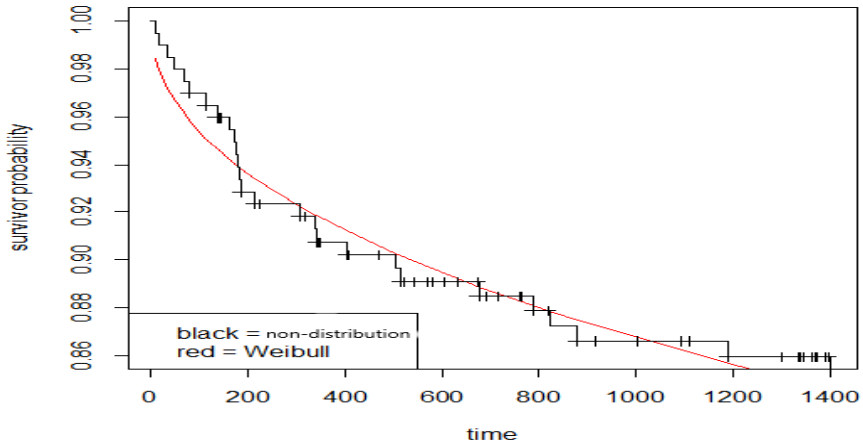

Figure 5: Weibull and non distribution survival curve for no predictor

Thus the Weibull model can now be estimated. Table 2 summarized the output of the Weibull model.

Table 2: Weibull output for no covariate

	Value	Std.Error	z
(Intercept)	11.330	0.819	13.83
log(scale)	0.721	0.182	3.97

$Scale = 2.06$

Weibull distributtion

Loglik(model) = -269.1

$loglik$(intercept only) $= -269.1$

Number of Newton-Raphson Iteration: 8

n= 199

So the survival and hazard functions corresponding to a prostate cancer patient with no covariate are:

$$\begin{aligned} S(t) &= \exp(-\exp(-11.33)^{2.06} t^{2.02}) \\ &= \exp(-731 \times 10^{-11} t^{2.06}) \end{aligned} \tag{29}$$

$$\begin{aligned} h(t) &= \exp(-11.3)^{2.06} 2.06 t^{1.06} \\ &= 15.06 \times 10^{-11} t^{1.06} \end{aligned} \tag{30}$$

It can been seen that the hazard rate of a prostate cancer patient increases with time. This means that the longer a patient stay in the study without any influence by variable, the higher his chance of dieing. The Weibull model gives a close approximation to the non parametric estimator by comparison. The result also shows that a prostate cancer patient will eventually experience death at the long run.

Modeling technique give a future estimate. For instance, we can estimate the survival and hazard time for a patient in the next 5000 days. The model only explain the survival and hazard rate of a patient but does not account for the factors that yield this results. It is therefore important to investigate the behavior of our model by taking into consideration the variable of the study (cancer stage) .

The AICs was used to fit which model fit best.

A graphical approach is use to assess the absolute goodness of fit of the Weibull model.

The table below shows the results of the AICs of the of the models under study. It look reasonable to still use the Weibull distribution to model the survival and hazard function for this study.

Table 3: AIC results for full variables

	Exponential	Log - Logistics	Weibull
AIC	558.6118	537.3886	537.3602

If the stage of the cancer of a patient was significant it must satisfy the Weibull assumption. This was done by plotting log(-log) of group three survivors against the log of the survival time. The same procedure is done for group four. The log-log for both stage one and stage two cannot be plotted and so have been omitted. This is because no patient experiences death for stage one cancer and so its log-log graph

cannot be estimated. Only one patient experiences the event and so its log-log graph is also omitted. Figure 6 provided the graph of log(-log) of survival against log of survival time. It can be observed that both graphs give an approximated straight line, suggesting that the Weibull assumption is met.

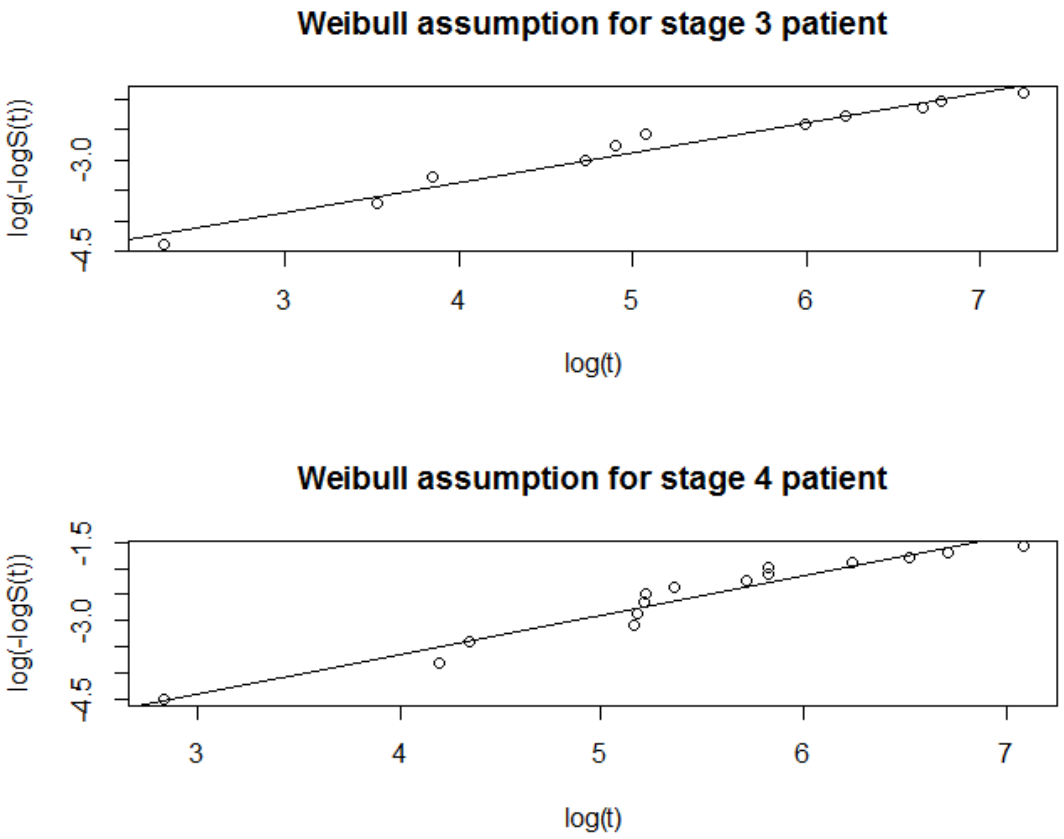

Figure 6: Weibull assumption for prostate cancer stages

Figure 7 provides the Weibull survival graph for cancer stages. The survival curve decreases downwards with time. Patients least survivors was estimated to be about 1%, 94%, 89% and 83% for stage1, stage2, stage3 and stage4 respectively.

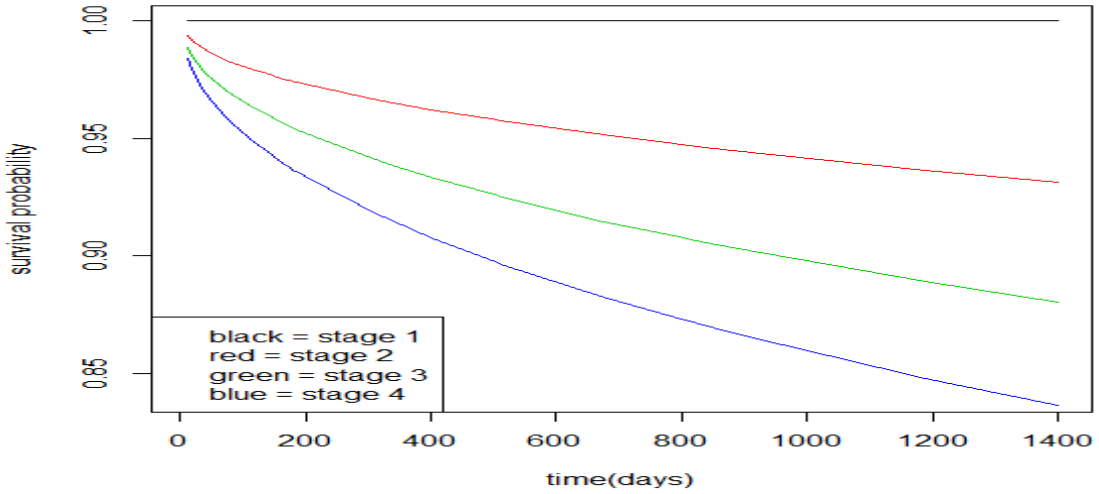

Figure 7: Weibull survival graph for prostate cancer stages

Figure 8 below shows the hazard graph for cancer stages. The hazard is observe to increase with time. Stage 4 and stage 3 curve are fairly close to each other but wide apart from stage 2 cancer. Thus stage 4 cancer was closer to stage 3 showing that both have similar hazard rate.

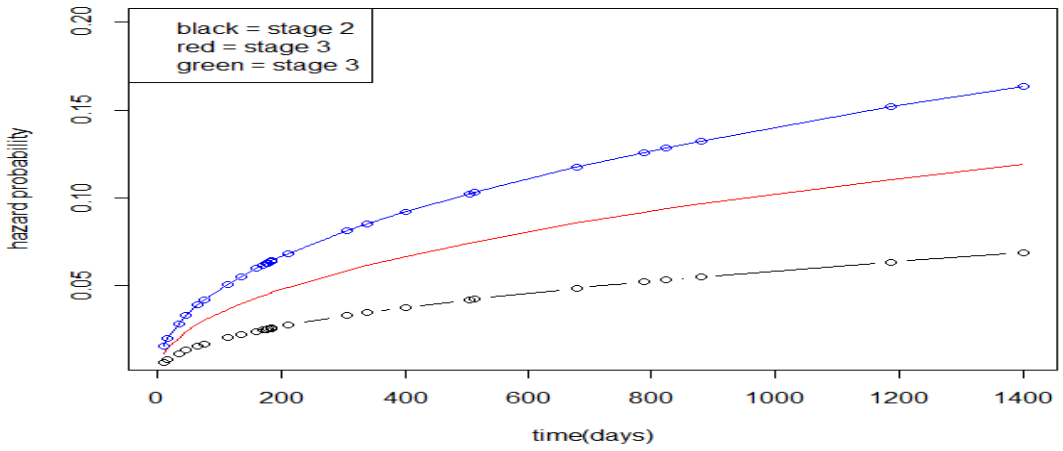

Figure 8: Weibull hazard for cancer stages

The estimated final model to predict survival and hazard functions for a prostate cancer patient is produce. Table 4 provided the output of the estimated Weibull.

Table 4: Weibull estimation

	Value	Std.Error	Z
(Intercept)	47.062	0.823	57.20
stage 2	-34.424	2.126	-16.19
stage 3	-35.614	0.815	-43.72
stage 4	-36.305	0.000	-Inf
Log(scale)	0.713	0.181	3.93

Scale =2.04

Weibull distribution

Loglik(model) = -266.6 Loglik(intercept only)= -269.1

$Numberof Newton - RaphonIteration : 20, n = 199$

There are two parameters in this output to consider. The Scale parameter and the intercept parameter. The Scale is K and the intercept is α in the following equations:

$$S(t) = \exp(-\exp(-\alpha)^k t^k)$$
$$H(t) = \exp(-\alpha)^k t^{k-1} \tag{31}$$

Thus, the survival and hazard function for this model are summarize in the table below:

Table 5: Weibull survival and hazard model

Survival stage model	Hazard stage model
$S(t, x_{stage1}) = \exp(-20.18 \times 10^{-43} t^{2.04})$	$h(t, x_{stag1}) = 41.17 \times 10^{-43} t^{1.04}$
$S(x_{stage2}) = \exp(-63.57 \times 10^{-13} t^{2.04})$	$h(t, x_{stag2}) = 12.97 \times 10^{-12} t^{1.04}$
$S(x_{stage3}) = \exp(-7.20 \times 10^{-11} t^{2.04})$	$h(t, x_{stag3}) = 14.69 \times 10^{-11} t^{1.04}$
$S(x_{stage4}) = \exp(-29.49 \times 10^{-11} t^{2.04})$	$h(t, x_{stag4}) = 60.16 \times 10^{-11} t^{1.04}$

The estimated acceleration factor γ comparing the cancer stage 2, stage 3 and stage 4 to stage 1: [(stage 1 vs. stage 2),(stage 1 vs. stage 3),(stage 1, stage 4)] is given by: $\gamma = exp(\alpha_i)$, where α_i is the coefficient of each cancer stage.

Table 6: Estimated acceleration factor for Patient stages

Cancer stage	$\gamma = exp(\alpha_i)$
stage 2	$1.122013e^{-15}$
stage 3	$3.413288e^{-16}$
stage 4	$1.710208e^{-16}$

Table 6 provided the acceleration factor for each cancer stage. The interpretation is as follows: The estimated coefficient for stage 2, stage 3 and stage 4 were negatives implying that the survival for patient with prostate cancer stage 2, stage 3

and stage 4 is decrease by a factor $of 1.122013e^{-15}$, $3.413288e^{-16}$ and $1.710208e^{-16}$ respectively. In all, the survival rate decreases with increase in prostate cancer stages.

Alternatively, the corresponding hazard can have its interpretation as having a prostate cancer stage 2 hastens death by a factor of $e^{34.42} = 88.80 \times 10^{15}$ relative to a stage 1 prostate cancer while having a prostate cancer stage 3 hastens death by a factor of $e^{35.61} = 29.19 \times 10^{16}$ relative to a stage 1 prostate cancer.

Having a prostate cancer stage 4 hastens death by a factor of $e^{36.31} = 58.78 \times 10^{16}$ relative to a stage 1 prostate cancer. Overall, the hazard of the stage of the cancer is seen to increase in stages relative to stage 1.

For instance, the hazard or risk of a patient dying with a prostate cancer in the next 15 years (5475 days) with stage 1, stage 2, stage 3 and stage 4 are $3.18 \times 10^{-38}, 10 \times 10^{-8}, 1.13 \times 10^{-6}and4.65 \times 10^{-6}$ respectively.

1 Appendix

1.1 R Code

```
***testing which parametric model to use for survival time with no covariate


weib=survreg(Surv(timee,vvent)~1,dist="w")

expp=survreg(Surv(timee,vvent)~1,dist="exponential")

loglog=survreg(Surv(timee,vvent)~1,dist="logl")

logm=survreg(Surv(timee,vvent)~1,dist="logn")


****** AIC for no each model with no predictor*

extractAIC(weib)[2]

extractAIC(expp)[2]

extractAIC(loglog)[2]

extractAIC(logm)[2]
```

```
**** Checking Weibull assumption for no predictor

kp=survfit(Surv(timee,vvent)~1)

summ=summary(kp)

tim=summ$time

suv=sum$surv

risk=summ$n.risk

lgtym=log(tim)

lgsv=log(-log(suv))

plot(lgtym,lgsv,xlab="log(t)",ylab=expression(log(-log(hat(S)*"(t)"))))

abline(lm(lgsv ~ lgtym))



************  weibull survivor curve for no predictor

plot(tim,1-pweibull(tim,1/2.1,exp(11)),ylim=c(0.86,1),xlab="time",

ylab="survivor␣probability",type="l")



********* weibull hazard curve for no predictor


plot(tim,pweibull(tim,1/2.1,exp(11)),ylim=c(0.02,0.3),xlab="time",

ylab="survivor␣probability",type="l")
```

```
**** ploting the weibul survivor curve and kaplan meier for no covariate

kp=survfit(Surv(timee,vvent)~1)

plot(tim,1-pweibull(tim,1/2.1,exp(11)),ylim=c(0.86,1),xlab="time"

,ylab="survivor␣probability",type="l",col="red")

lines(kp,conf.int=FALSE)

legend("bottomleft",legend=c("black␣=␣kaplan␣Meir","red=Weibull"),col=1:2)
```

```
********* Estimating weibull for no predictor

w=survreg(Surv(timee,vvent) ~1,dist="weibull",scale=0)

summary(w)
```

```
**********   AIC selection for full model
```

```
wete=survreg(Surv(timee,vvent)~gg+as.factor(tmtt)+as.factor(sstg),dist="w")


expte = survreg(Surv(timee,vvent) ~ gg+as.factor(tmtt)+as.factor(sstg),
 dist="exponential")


 logte=survreg(Surv(timee,vvent)~gg+as.factor(tmtt)+as.factor(sstg),
 dist="logl")
 extractAIC(wete)[2]
 extractAIC(expte)[2]
 extractAIC(logte)[2]



********* checking weibull assumption for cancer stage
stage=survfit(Surv(timee,vvent) ~as.factor(sstg))
sum=summary(stage)
suv3=survfit(Surv(timee[sstg==3],vvent[sstg==3])~ 1)
suv4=survfit(Surv(timee[sstg==4],vvent[sstg==4]) ~ 1)
sum3=summary(suv3)
sum4=summary(suv4)
suvstg3=sum3$surv
time3=sum3$time
suvstg4=sum4$surv
```

```
time4=sum4$time

ti=sum$time

tim3log=log(time3)

suvstg3log=log(-log(suvstg3))

tim4log=log(time4)

suvtg4log=log(-log(suvstg4))

par(mfrow=c(2,1))

plot(tim3log,suvstg3log,xlab="log(t)",ylab="log(-logS(t))")

abline(lm(suvstg3log ~ tim3log))

plot(tim4log,suvtg4log,xlab="log(t)",ylab="log(-logS(t))")

abline(lm(suvtg4log ~tim4log))
```

```
*******  ploting the weibull survival for cancer stages
```

```
plot(tim,1-pweibull(tim,1/2.04,exp(47.062)),xlab="time(days)",

ylab="survivor probability",ylim=c(0.84,1),type="l")

lines(tim,1-pweibull(tim,1/2.04,exp(12.638)),col=2)

lines(tim,1-pweibull(tim,1/2.04,exp(11.448)),col=3)
```

```
lines(tim,1-pweibull(tim,1/2.04,exp(10.757)),col=4)
legend("bottomleft",legend=c("black = stage 1","red = stage 2",
 "green = stage 3", "blue = stage 4"),col=1:2)


********  ploting the weibull hazard for cancer stages

plot(tim,pweibull(tim,1/2.04,exp(12.6)),ylim=c(0.01,0.2),xlab="time(days)",
ylab="hazard probability",type="b")
lines(tim,pweibull(tim,1/2.04,exp(11.)),type="l",col=2)
lines(tim,pweibull(tim,1/2.04,exp(10.4)),type="o",col=4)
legend("topleft",legend=c("black = stage 2","red = stage 3", "green =
 stage 3"),col=1:3)


*********fitting Weibull model***********

weee=survreg(Surv(timee,vvent)~as.factor(sstg),dist="w")
summary(weee)
```