COMPARISON OF COX, WEIBULL AND GOMPERTZ REGRESSION MODELS IN SURVIVAL ANALYSIS USING BREAST CANCER DATA

BY

USMAN MOHAMMED
MSC/SCIEN/12360/2011 – 2012

DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
AHMADU BELLO UNIVERSITY, ZARIA – NIGERIA

NOVEMBER, 2014

COMPARISON OF COX, WEIBULL AND GOMPERTZ REGRESSION MODELS IN SURVIVAL ANALYSIS USING BREAST CANCER DATA

BY

USMAN MOHAMMED

MSC/SCIEN/12360/2011 – 2012

BEING

A THESIS SUBMITTED TO THE POST GRADUATE SCHOOL, AHMADU BELLO UNIVERSITY, ZARIA, DEPARTMENT OF MATHEMATICS, IN PARTIAL FULFILMENT FOR THE AWARD OF DEGREE OF MASTER OF SCIENCE (M.SC.) IN STATISTICS.

NOVEMBER, 2014

DECLARATION

I declare that the work in this thesis entitled "COMPARISON OF COX, WEIBULL AND GOMPERTZ REGRESSION MODELS IN SURVIVAL ANALYSIS USING BREAST CANCER DATA" has been performed by me in the Department of Mathematics under the supervision of Prof. S. U. Gulumbe and Dr. H.G. Dikko. The information derived from the literature has been acknowledged in the text and a list of references provided. No part of this thesis was previously presented for another degree or diploma at any university.

Usman Mohammed

| --- | --- | --- |
| Name of student | Signature | Date |

CERTIFICATION

This thesis entitled "COMPARISON OF COX, WEIBULL AND GOMPERTZ REGRESSION MODELS IN SURVIVAL ANALYSIS USING BREAST CANCER DATA" by Usman Mohammed meets the regulations governing the award of the degree of Master of Science of Ahmadu Bello University, and is approved for its contribution to knowledge and literary presentation.

---------- ----------

Prof. S. U. Gulumbe Date
Chairman, Supervisory Committee

---------- ----------

Dr. H. G. Dikko Date
Member, Supervisory Committee

…….. ……

Dr. W. B. Yahya Date
External Examiner

---------- ----------

Dr. B. Sani Date
Head of Department

---------- ----------

Prof. A. Z. Hassan Date
Dean Postgraduate School

AKNOWLEDGEMENT

I am grateful to almighty Allah (S.W. A) for granting me the opportunity to undertake this thesis work, from the beginning up to the end.

First, my special thanks and appreciation go to my supervisors. Prof. S.U. Gulumbe and Dr. H. G. Dikko for their guidance and invaluable advice during the research, and for their contributions and criticism in the preparation of this thesis. My appreciation also goes to Mal. Shehu Bala for his contribution towards the success of this research. I'm indebted to Dr. Alfred A. Abiodun of Dept. of Statistics, University of Ilorin in the area of data analysis, your contribution form part of this academic attainment. I wish to express my sincere gratitude to Mr. John A. Obemeata of Data Processing Unit of Institute for Agricultural Research, Ahmadu Bello University, Zaria, my colleagues in M.Sc. class 2012 and all the academic and non academic staff of the Department of Mathematics, Ahmadu Bello University, Zaria for their invaluable contributions towards the success of this research work.

I wish to extend my gratitude to the management of Ahmadu Bello University Teaching Hospital (ABUTH), Zaria for supplying the necessary data for the research work. Finally, I wish to thank my family for their prayer and encouragement throughout the period of this research. I wish all of us peace, happiness and success in all what we do in life.

ABSTRACT

Survival analysis is a class of statistical methods designed to study the occurrence and timing of events. This study aimed to compare the results of semi-parametric Cox proportional hazards model and parametric models (Weibull and Gompertz) to determine the model that best fits breast cancer data. Kaplan-Meier (K-M) method was used to estimate and graph the survival curves using the data obtained from Ahmadu Bello University Teaching Hospital Zaria on 312 breast cancer patients between 1997 and 2012. The mean age of the breast cancer patients was found to be 43.39 years with standard deviation of 11.74 years and overall median survival time of 10 months. The 5-year overall survival rate was about 35.0%. In comparing the semi-parametric Cox model and parametric (Weibull and Compertz) models, Akaike Information Criterion (AIC) was used to evaluate the three models. Weibull regression model had the least AIC value (422.60) which shows best performance in handling breast cancer data, where as Cox regression model has the highest AIC value (530.65) followed by Gompertz model with AIC value (430.28). From the results of the analysis obtained, for Cox, Weibull and Gompertz regression models, age, occupation and stage II of the breast cancer does not have significant effect on the mortality of the patients, ($p = 0.0440, 0.0270, 0.1740$ respectively) but results of the treatment and stage III of breast cancer have significant effect on the mortality of the patients, ($p = 0.0001, 0.00001$ respectively). $p < 0.01$ is considered as statistical significant. The results of this study showed that, according to our breast cancer data, the parametric Weibull regression model could better determine the factors associated with the breast cancer disease than the semi-parametric Cox proportional hazards model. That is, Weibull

model provided a better fit to the study data than the Cox proportional hazards and Gompertz models.

TABLE OF CONTENTS

Cover page……i

Title page……..ii

Declaration …….....iii

Certification ……iv

Acknowledgement ……...v

Abstract ……..vi

Tale of contents ……vii

List of Tables …….....x

List of Figures…….xi

CHAPTER ONE: BACKGROUND TO THE STUDY

1.1 Introduction…….1

1.2 Statement of the problem……5

1.3 Aim of the Study……..5

1.4 Objectives of the study……5

1.5 History of Cancer Disease……5

1.5.1 Causes of Cancer Disease…….6

1.5.2 Those at Risk of Cancer……...7

1.5.3 Types of Cancer……7

1.5.4 Ways to Reduce the Risk of Cancer……7

1.5.5 Diagnosis of Cancer …….....8

1.5.6 Treatment of Cancer ……...9

1.6 Breast Cancer…….11
1.7 Basic Concepts in Survival Analysis……..12

1.7.1 Survival Function……..…....12

1.7.2 Hazard Function……13

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction……15

2.2 Non Parametric Approach…….....17

2.3 Comparison of Survival Distribution…….22

2.4 Semi-Parametric Approach ……...25

2.5 Parametric Approach……..30

2.5.1 Weibull Regression Model…….30

2.5.2 Gompertz Regression Model……..32

CHAPTER THREE: METHODOLOGY

3.1 Introduction ……...35

3.2 Kaplan-Meier Estimator…….....35

3.3 Comparison of the Survival Curves of Two Groups Using the Log-rank Test……..37

3.4 Cox Proportional Hazard Model……38

3.5 Cox Proportional Hazard Model Partial Likelihood Estimation……39

3.6 Weibull Regression Model…….41

3.7 Gompertz Regression Model……..43

3.8 Akaike Information Criterion (AIC)……..46

CHAPTER FOUR: ANALYSIS AND DISCUSSION OF RESULTS

4.1 Introduction……47

4.2 Kaplan –Meier (K-M) Survival Function……...47

4.3 Log- rank Tests for Equality of Survival Functions (Survival Probabilities)…..…50

4.4 Test for Proportional Hazard Assumption……..53

4.5 Results from the Analysis of Cox Proportional Hazard Model……..57

4.6 Results from the Analysis of Weibull Model…….59

4.7 Results from the Analysis of Gompertz Model……..61

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary……65

5.2 Conclusion……..66

5. 3 Recommendations……..68

5.4 Suggestions for Further Research……..68

References …….69

Appendix I……..77

LIST OF TABLES

Table 4.1: Stages of the Breast Cancer Patients……50

Table 4.2: Occupations of the Breast Cancer Patients……..51

Table 4.3: Age groups of the Breast Cancer Patients…….....51

Table 4.4: Results of the Treatment of the Breast Cancer Patients……52

Table 4.5: The Summary of Log-rank Tests for Group Variables……52

Table 4.6: Results for Cox Model…….57

Table 4.7: Results for Weibull Model……59

Table 4.8: Results for Gompertz Model…….61

Table 4.9: Summary of the Results of the Analysis and Akaike Information Criterion (AIC) for Cox, Weibull and Gompertrz Regression Models……...63

LIST OF FIGURES

Figure 2.1: Survival Function by K. - M. Estimator……19

Figure 4.1: Kaplan -Meier (K-M) curves for overall survival estimate……..…48

Figure 4.2: Kaplan-Meier (K-M) Curves for survival estimate of occupation……..49

Figure 4.3: Kaplan-Meier (K-M) Curves for survival estimate of age……49

Figure 4.4: Test of Proportional Hazard Assumption for Age Groups……53

Figure 4.5: Test of Proportional Hazard Assumption for Stages of Breast Cancer…….54

Figure 4.6: Test of Proportional Hazard Assumption for Results of the Treatment……55

Figure 4.7: Test of Proportional Hazard Assumption for Occupation of the Patients…….56

CHAPTER ONE

1.1 Introduction

Survival analysis is a class of statistical methods designed to study the occurrence and timing of events. The methods were developed for studying the time from initiating events (such as birth, diagnosis of a disease, start of a treatment e.t.c.) to some terminal events (such as response to a treatment, relapse, death e.t.c.). These methods are mostly applied in medical sciences. However, the methods were successfully applied to many different kinds of events across disciplines. Examples include the manufacturing or engineering sectors (how long it takes a machine to fail), social sciences (how long a marriage will last), finance (the time it takes the price of stock in stocks market to drops) and so on. Sometimes other names are used to refer to this class of methods such as "duration analysis", "failure time analysis", "event history analysis", "transition analysis", but the basic techniques are the same as in the underlying ideas i.e. understanding the pattern of events in time and what factors cause those events to occur.

Survival analysis measures the probability of how long it takes a given outcome to occur for a group of similar individuals known as failure times (Cox and Oates, 1984). Survival analysis method also measures the probability that the given event in question will last the next point in time (Gokovali *et al,* 2007). Survival analysis studies are well known to occur in medical sciences particularly in cancer studies. Depending on the question of the studies, one is interested in the risk groups i.e. which group of patients or components is more likely to experience the events; or time predictions. The starting time from a defined point to the occurrence of a given event is called survival time (Berwick *et al, 2004*) and the analysis of such group data is called survival analysis (Altman, 1977).

In order to understand the survival analysis; the concept of event has to be understood. According to Blossfeld *et al,* (1989) an event is a change in state as defined by one or more qualitative variables within some observational period and within the relevant state space. Examples are death, marriage, promotion e.t.c. Survival analysis is the core of any study of time to a particular event, such as death, infection or diagnosis of a particular disease. It is therefore fundamental to most epidemiological cohort studies. Survival data can be thought of as being generated by what is called a "failure time process". A failure time process consist of units, such as individuals, government, countries and so on that are observed at some starting point in time. These units are in some states. For example, the individual is healthy, the government is in power, a country is at peace and so on, and is observed over time. At any given point in time, these units are at 'risk' of experiencing some events, where an event essentially represents a change or transition to another state. For example, the individual dies, the government falls from power, a country is at war e.t.c. After the event is experienced, the unit is either no longer observed or it is at risk of experiencing another kind of event.

In some circumstances, units are not observed in experiencing an event; that is no transition is made from one state to another while the unit is being observed. For example, the individual remains healthy, the government remains in power, the country remains at peace and so on. These cases are called 'censored' since we do not observe the subsequent history of the unit after the last observation point. Censoring is endemic to survival analysis data. It is defined as the loss of observation on the life time variable of interest in the process of an investigation. In survival data, censoring frequently occurs for many reasons. In clinical trial, the effectiveness of a new medical treatment for a particular disease is

investigated. A patient for example, may be lost to follow up due to migration or health problems. Censoring occurs when a subject does not have an event during the observation time, meaning that we cannot observe what has happened subsequently on the subject. According to Klienbaum and Klien (2005), there are generally three reasons why censoring may occur; when a person does not experience the event before the study ends; when a person is lost to follow-up during the study period; when a person withdraws from the study because of death (if death is not the event of interest) or some other reason.

There are different types of censoring. They are right censoring, left censoring and interval censoring. Right censoring is the most common type of censoring, it occurs when the observation is terminated. In other words, it occurs when a subject leaves the study before an event occurs. For example, consider the patients in a clinical trial for the study of the effect of treatments on stroke occurrence. The study ends after five years, those patients who have no strokes by the end of the five years are said to be right censored. Also, in a study of divorce, couples that do not divorce during the observational period are right censored. Left censoring rarely occur, it happens when an event is known to have occurred before some particular time, but the exact time is unknown. For example, infection with some sexually transmitted disease such as HIV/AIDS.

Another example is in a study of first marriage, if a person is known only to have married before age 20, that person's marriage age is left censored. For interval censoring, if an individual or subject is known to have experienced an event within an interval of time but the actual survival time is not known, then an interval censoring is said to have occurred. The actual occurrence time of an event is known within an interval of time. For example, infection with sexually – transmitted diseases such as HIV/AIDS with regular testing e.g. annually.

Another feature of survival data is" Truncation". Truncation refers to complete lack of information about the occurrence of the event. It is a condition other than the event of interest that is, for example, used to screen respondents or patients (Klein and Moeschberger, 1997). There are three types of truncation. They are right, left and interval truncation. The most common type of truncation is left-truncation, this occurs when subjects enter the study at a random age. In the case of left truncation, we do not have information before the onset of risk to sometime after the risk. In other words, the subject was not observed for some time at the beginning of the process but then came under observation.

Right-truncation occurs, but is less frequent. Klein and Moeschberger (1997) provided the example of the examination of an episode from HIV infection until the development of AIDS. If the sample only includes those who have developed AIDS prior to the end of the study, those HIV-infected individuals who have not yet progressed to AIDS are excluded from the sample. Another type of truncation is interval or gap truncation, which is similar to left-truncation. This could occur in a clinical study if, for example, a patient is under observation for the first 3 months of the study, drops out for 2 months and then rejoins the study again for the last 7 months. Dropping out of the study for 2 months creates an interval or gap in the period of observation. Both left and interval truncations are dealt with during analyses by omitting the subject from all individual in the study (in a binary-outcome analysis) during the truncation period (or gap) due to the fact that they could not have experienced the event during those periods.

1.2 Statement of the problem

Cox Proportional Hazard Model is a semi-parametric model, whereas Weibull and Gompertz models are parametric models. These models are mostly used in the analysis of survival data in clinical research. It has not been established which model among the three that is most efficient in the analysis of survival data. This study therefore, compared the three models towards establishing the one that would be most suitable for the analysis of survival data in clinical research.

1.3 Aim of the Study

Cancer disease differs, as the models used for cancer studies also differs. This study aim to investigate semi-parametric Cox proportional hazard model and parametric (Weibull and Gompertz) models in assessing the best comparative model that fits breast cancer data in clinical research.

1.4 Objectives of the study

The following are the objectives of the study:

-To describe the survival function using Kaplan-Meier (K-M) approach, and compare the survival curves using Log-rank tests.

-To fit the three models used in the survival analysis using data on breast cancer.

-To evaluate the models used in the study of survival analysis using model comparison.

1.5 History of Cancer Disease

Human cancer is probably as old as human race. The word cancer came from the father of medicine, Hippocrates a Greek physician. Hippocrates used the Greek words, ‘carcinos’ and ‘carcinoma’ to describe tumour thus, calling cancer “karkinos”. These Greek

terms actually were words that describe a “crab”, which Hippocrates thought a tumour resembled. Tumour is abnormal mass of new tissue growing in or on the body, causing cancer disease; it destroys parts of the body, often causing death. Although, Hippocrates was certainly not the first to discover the disease.

The history of cancer began much earlier. Cancer is one of the ten causes of death of the world population. World Health Organisation (WHO)'s International Agency for Research on Cancer also found the number of cancer-related deaths increased from 7.6 million in 2008 to 8.2 million in 2012. The most commonly diagnosed types of cancer were lung cancer, breast cancer and colorectal cancer, which made up 13 percent, 11.9 percent and 9.7 percent of the total, respectively

1.5.1 Causes of Cancer Disease

Normally, the body makes new cells when they are needed for the growth or repair of damage parts of the body. This cells growth is controlled by a group of genes that works together. When one or more of these genes is damaged (or mutated), too many cells are produced. These excess cells clump together to form a tumour resulting in cancer. Genes can be mutated by bad chemicals in a person’s environment , like tobacco smoke, but more often, damage to growth control genes occurs spontaneously in only a few cells of the body, that is why a tumour is typically found or at least begins in just one spot.

Cancer is not just one disease, but a large group of almost 100 diseases. Its two main characteristics are uncontrolled growth of cells in the human body and the ability of

these cells to migrate from original site and spread to distant sites. If the spread is not controlled, cancer can result in death.

1.5.2 Those at Risk of Cancer

Most people who developed cancer have no family history of the disease, meaning that their cancer was not inherited, but sometimes the more common cancer (example breast, prostate and colon cancer) can run in a family. If a close relative, (example, grand parent, parent or sibling) has been diagnosed with a common form of cancer, you too may be at risk. Your risk increases if a relative develop the cancer at an early age i.e. before age 50 – 55 years, or if more than one close relative develop the same cancer.

1.5.3 Types of Cancer

Cancer has the potential to affect every organ in the body. The cells within malignant tumour have the ability to invade neighbouring tissues and organs, thus, spreading the disease. It is also possible for the cancerous cells to break free from the tumour and enter the blood stream, in turn spreading the disease to other organs. This process of spreading is called metastasis.

Some of the different types of cancer are:

Blood cancer, Bone cancer, Brain cancer, Breast cancer, Digestive/Gastrointestinal cancer, Endocrine cancer, Respiratory cancer, Skin cancer, Prostate cancer, Cervical cancer

1.5.4 Ways to Reduce the Risk of Cancer

The best way to reduce cancer risk may differ slightly for each type of cancer. Here are some general guidelines to follow:

-Do not use tobacco products; avoid excessive alcohol consumption (for lung cancer)

- perform regular self-examination.

-See your doctor for periodic screening.

-If cancer runs in your family, discuss this information with your doctor.

-Maintain a healthy body weight, avoid obesity.

-Exercise regularly.

-Eat plenty of vegetables and fruits.

-Decrease or avoid consumption of animal fats and red meat.

1.5.5 Diagnosis of Cancer

Diagnosis begins with physical examination and a complete medical history. The doctor observe, feel and palpate (apply pressure by touch) different parts of the body in order to identify any variations from the normal size. As part of the physical examination, the doctor will inspect the oral cavity, or the mouth. To detect thyroid cancer, the doctor will observe the front of the neck for swelling.

The doctor may conduct a thorough examination of the skin to look for sores that have been present for more than three weeks. Examination of the female pelvis is used to detect cancers of the ovaries, uterus, cervix and vagina. Breast examination includes visual observation where the doctor looks for any discharge, unevenness, disclouration or scaling, he palpates both breasts to feel for masses or lumps.

For males inspection of the rectum and the prostate is also included in the physical examination. The doctor inserts a gloved finger in to the rectum and rotates it slowly to feel for any growths, tumours or other abnormalities. If the doctor detects an abnormality on

physical examination, or the patient has some symptom that could be indicative of cancer, the doctor may order diagnostic tests.

A blood test for cancer is easy to perform, usually inexpensive and risk-free. Blood tests can be either specific or non-specific. Often, in certain cancers, the cancer cells release particular proteins (called tumour makers) and blood tests can be used to detect the presence of these tumour makers. Blood tests are generally more useful in monitoring the effectiveness of the treatment, or in following the course of the disease and detecting re-occurrence of the disease.

1.5.6 Treatment of Cancer

Treatment and prevention of cancers continue to be the focus of a great deal in cancer research. Research in to new cancer therapies included cancer targeting gene therapy, virus therapy and a drug that stimulate apoptosis or self destruction of cancer cells, but not healthy cells. However, all of these new therapies take years of clinical testing and research.

The aim of cancer treatment is to remove all or reduce as much of the tumour as possible and prevent the recurrence or spread of the primary tumour. If the cancer is very aggressive and a cure is not possible, then the treatment should be aimed at relieving symptoms and controlling the cancer for as long as possible. Cancer treatment can take many different forms, and it is always tailored to the individual patient. The decision of which the type of treatment is the most appropriate depends on the type and location of cancer, the extent to which it has already spread the patient's age, sex, general health status and personal treatment preferences. The major types of treatment are:

Surgery, radiation, chemotherapy, immunotherapy, hormone therapy and bone-marrow transplantation.

Surgery: Surgery is the removal of a visible tumour and is the most frequently used cancer treatment. It is most effective when a cancer is small and confined to one area of the body.

Radiation: Radiation kills tumour cells. Radiation is used alone in cases where tumour is unsuitable for surgery. Most often, it is used in conjunction with surgery and chemotherapy. Radiation can either be external or internal. In the external form of radiation, the radiation is aimed at the tumour from outside the body. In internal radiation, (also known as brachytherapy), a radioactive substance in the form of pellets or liquid is paced at the cancerous site by means of a pill, injection or insertion in a sealed container.

Chemotherapy: This is the use of drugs to kill cancer cells. It destroys the hard-to-detect cancer cells that have spread and are circulating in the body.

Immunotherapy: This uses the body's own immune system to destroy cancer cells. This form of treatment is being intensively studied in clinical trials and is not widely available to most cancer patients.

Hormone Therapy: Hormone Therapy is a standard treatment for some types of cancer that are hormone-dependent and grow faster in the presence of a particular hormone. These include cancer of the prostate, breast and uterus. Hormone Therapy involves blocking the production or action of these hormones. As a result, the growth of the tumour slow down and survival may be extended for several months or years.

Bone marrow transplantation: The bone marrow is the tissue within the bone cavities that contains blood-forming cells. Healthy bone marrow tissue replenishes the blood supply and is essential to life. Sometimes the amount of drugs or radiation needed to destroy cancer cells also destroys bone marrow. Replacing bone marrow with healthy cells

counteracts this adverse effect. A bone marrow transplant is the removal of marrow from one person and the transplant of the blood-forming cells either to the same person or to someone else.

1.6 Breast Cancer

Breast cancer is a type of cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk (Sariego, 2010). Cancers originating from ducts are known as ductal carcinomas, while those originating from lobules are known as lobular carcinomas. Breast cancer occurs in humans and other mammals. While the overwhelming majority of human cases occur in women, male breast cancer can also occur. Breast cancer is the most frequent cancer in women while it is the second one in all humans. Although its prevalence changes in different societies, it is known that one out of 8-10 women in Western society would have breast cancer during their life span.

The mortality rate of breast cancer was calculated as 3%. The prevalence of the disease is especially higher in the North America and European countries than the rest of the world (Stewart and Kleihues, 2003). In 1990, the World Health Organization (WHO) conducted a study in which 796,000 breast cancer cases were reported out of which 314,000 deaths occurred. In 2002, 1,152,000 new cases and 411,000 deaths were detected in a study, conducted by International Agency on Cancer for Research (IARC) under WHO administration. According to all phases of the diseases, the five year survival rate has been reported as 73% in developed countries and 52% in developing countries. The significant difference between those two values can be explained by early diagnosis via

mammography scans and better treatment in developed countries. Breast cancer is a disease found mostly in females, and it is a genetic transmission disease, causing by abnormal hormones, obesity and food with high fats and spreading of cancer from other part of organs to the breast (Wingo, *et al* 1995). Therefore breast cancer is considered as a major health issue worldwide (Harirchi, *et al* 2001).

1.7 Basic Concepts in Survival Analysis

The primary concept in survival analysis is the survival time which is also the failure time. Survival time is the length of time that is measured from time origin to the time the event of interest occurred. Other concepts are survival and hazard functions.

1.7.1 Survival function

A typical analysis of survival data involves the modeling of time-to- event data, such as time until death. The time to event of interest is called either survival time or failure time. Survival function is a basic quantity in nature, employed to describe the probability that an individual survives beyond a specific time. In other words, survival function is the amount of time until the event of interest occurs. Survival function gives information on the probability of a time-to-event of interest. The survival function, models the probability of an individual surviving beyond a specific time‘t’. The statistical expression of survival function is given by

$$S(t) = P(T \geq t), \quad t \geq 0 \tag{1.1}$$

‘ T ’ is denoted as random variable representing the survival time, which is the time until the event of interest occurred and ‘ t ‘ is any particular value of ‘ T ‘. In other words, the probability of experiencing the event of interest beyond time ‘t ‘is modeled by the survival function.

Note that survival function is defined for both discrete and continous T, and the probability density function and the hazard function are easily specified for discrete and continuous T.

For discrete random variable T, taking values $0 \le t_1 < t_2 < \ldots$, let the probability mass function be given by $P(T = t_i) = f(t_i); \quad i = 1, 2, \ldots$ then the survival function is

$$S(t) = \sum_{j/t_i \ge t} f(t_j) \tag{1.2}$$

$$= \sum f(t_j) I(t_j \ge t) \tag{1.3}$$

where

the indicator function $I(t_j \ge t) =$ $\begin{cases} 0, & \text{if } t_j > t \\ 1, & \text{if } t_j \le t \end{cases}$ (1. 4)

For continuous variable T, the probability density function of T is

$$f(t) = F'(t) = -S'(t); \qquad t \ge 0 \tag{1.5}$$

1.8.2 **Hazard function**

The hazard function gives the instantaneous failure rate at time‘t ’ given that the individual has survived up to time ‘t’ i.e.

The hazard function $h(t) \ge 0$, is given as

$$h(t) = \lim_{\Delta t \to 0} \frac{p \ \ individual \ \ \ fails \ \text{in} \ \ time \ \ \text{int}erval(t, t + \Delta t) / it \ \ survives \ \ until \ \ time \ 't'}{\Delta t} \tag{1.6}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t / T \ge t)}{\Delta t}; \qquad t \ge 0 \tag{1.7}$$

where

Δt = small interval of time

There is a defined relationship between S(t) and h(t) which is given by the formular

$$h(t) = \frac{f(t)}{S(t)} = \frac{-dLogS(t)}{dt} \tag{1. 8}$$

$$S(t) = \exp\left[-\int_0^t h(u)du\right] = \exp(-H(t)) \quad ; \quad t \ge 0 \tag{1.9}$$

where

$H(t) = \int_0^t h(u)du$ is called cumulative hazard function, which can be obtained from the survival function since h (t) = -log S(t) (1. 10)

The probability density function of T can be written as

$$f(t) = h(t)\exp\left[-\int_0^t h(u)du\right]; \qquad t \ge 0 \tag{1.11}$$

The survival function is most useful for comparing the survival progress of the two or more groups. The hazard function gives a more useful description of the risk of failure at any point time.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The most useful benefits to survival models are keen to their growth in popularity. First, survival models are able to take time-varying variables in to account within the modeling process (Golub, 2007). Secondly, survival models are not restricted to the assumption that the distributions of the variables in the data must be normal (Sloot and Verchuren, 1990). Thirdly, survival models only produce positive predictions of time (Gokovali, *et al* 2007). It follows that time has the potential not to follow normal distribution, and need to be positive in predictions and is influenced by time varying variables.

Survival analysis is a collection of statistical procedures for analysing data, for which the outcome variable of interest is the time until an event occurs. It is the study of time between entry into observation and a subsequent event. The term 'Survival analysis' came into being from initial studies, where the event of interest was death. The term 'event' is simply a transition from one discrete state to another at an instantaneous moment in time. Nowadays, the scope of the survival analysis has become wide. Today, scientists are using it for time until onset of disease, time until earthquake, financial experts are using it for time until stock market crashes, and engineers are using it for time until equipment fail, and so on. The most common events studied are death, disease, relapse, and recovery. Few examples of studies, where tools of survival analysis are used are: leukemia (blood cancer) patients and time in remission, time to develop a heart disease for normal individuals,
13-year follow up of an elderly population (60+ year) to see how subjects remain alive and time until death, of heart transplants. (Kleinbaum and Klien, 2005).

The time when the statistical procedure of survival analysis started was not known. Probably it originated centuries ago, but it was only after World War II that a new era of

survival analysis emerged, stimulated by an interest in the reliability of military equipment. At the end of the war, these newly developed statistical methods, emerging from strict mortality data research to failure time research, quickly spread through the private industry as customers became more demanding of safer and more reliable products. As the uses of survival analysis grew, parametric models gave way to nonparametric and semi parametric approaches for their appeal in dealing with the ever-growing field of clinical trials in medical research.

Survival analysis was well suited for such work because medical intervention follow-up studies could start without all experimental units enrolled at start of observation time and could end before all experimental units had experienced an event. This is extremely important because even in the best-developed studies, there will be subjects who choose to quit participation, who move too far away to follow, or who will die from some unrelated events. The researcher was no longer forced to withdraw the experimental unit and all associated data from the study; instead techniques called censoring enabled researchers to analyze incomplete data due to delayed entry or withdrawal from the study. This was important in allowing each experimental unit to contribute all of the information possible to the model for the amount of time the researcher was able to observe the unit. The last great stride in the application of survival analysis techniques has been a direct result of the availability of software packages and high performance computers which are

now able to run these difficult and computationally intensive algorithms relatively efficiently.

In medical research, the following models or techniques of survival analysis are mainly used to obtain the estimates of survival probabilities, mean or median survival time.

The method of survival analysis is also used for the comparison of two or more treatments or procedures. The three techniques are:

Non parametric, Semi parametric and Parametric

2.2 Non Parametric Approach

The simplest and easy method which is free from all assumptions underlying parametric test is 'non parametric'. The methods of non parametric are often easy and simple to understand as compared to parametric methods. Non parametric analyses are more widely used in situation where there is doubt about the exact form of distribution. The most popular and common used method of non parametric is the Kaplan – Meier (K-M) method or product limit method (Kaplan and Meier, 1958). Lee, (1982), recommended K–M estimator as a statistical technique for use in cancer trials. It is applied in analysing the distribution of the patient's survival times following their recruitment in the study. The analysis expresses this in terms of proportion of patient still alive up to a given time following the recruitment or entry to a study.

The K-M estimator is also called a non parametric maximum likelihood estimator. It is used for estimating the survival probabilities. The method computes the probability of dying at a certain point in time conditional to the survival up to that point. It utilises the information of censored individuals till the point when the patient is censored. Thus, it maximizes utilisation of available information on time to event of the study sample.

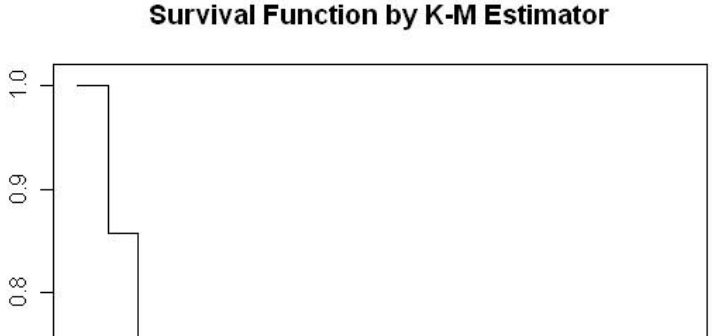
This is a modified form of the" Life Table" technique, with the condition that each time interval contains exactly one event and event occurs at the beginning of the time interval. In clinical studies, individual data are usually available on time to death or time to last seen alive. The life table technique is one of the oldest methods for analyzing survival data.

The distribution of survival is divided into a certain number of intervals. For each interval, the number and proportion of cases or objects that entered the respective interval alive is computed, the number and proportion of cases that failed in the respective interval (number of terminal events, or number of cases that died) and the number of cases that were lost or censored in the respective interval will also be computed. Based on those numbers and proportions, several additional statistics can be computed; such as number of cases at risk, proportion failing, proportion surviving, survival function hazard rate and median survival time. This procedure is used for large samples where the time intervals are large enough to be broken down in to smaller units. Van Groenendael (2010) applied K-M method to compare surgery and endogenous laser ablation in a retrospective study. The result of the study was time to reoccurrence rates under two interventions.

The K-M estimator for the survival curves is usually used to analyse individual data, whereas the life table method applies to group data. Since the life table method is a group data statistic, it is not as precise as the K-M estimate, which uses individual values.

A typical pattern of K-M survival curve is shown in figure 2.1

Figure 2.1



Suppose that n–individuals have failures in a group of individuals. Let $0 \le t_{(1)} < t_{(2)} <,...,< t_{(n)} < \infty$ be the observed ordered death times. Let n_j be the size of the risk set at $t_{(j)}$, where the risk set denotes the collection of individuals alive and uncensored just before $t_{(j)}$. Let d_j be the number of observed death at $t_{(j)}$ $j = 1, 2, ..., n$, then the K-M estimator of the $S(t)$ is defined by

$$\hat{S}(t) = \prod_{t_{(j)} < t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.1}$$

The estimator is a step function that changes values only at the time of each death. In fact K-M estimator will be shown to maximize the likelihood in the discrete case (Cox and Oakes, 1984).

Suppose that the distribution is discrete, with small quantity h_j at finitely many specified points, $0 \le \tau_1 < \tau_2 <,...,< \tau_j$ the survival function $S(t)$ may be expressed in terms of the discrete hazard function h_j as

$$S(t) = \prod_{\tau_j < t} (1 - h_j) \tag{2.2}$$

To derive the likelihood function from a sample of n observations, we first collect all the terms corresponding to the small quantity T_j. Let b_i = j if the i^{th} individual dies at T_j, then the contribution to the total log likelihood is

$$\log h_{b_i} + \sum_{k<b_i} \log(1 - h_k) \tag{2.3}$$

Let e_i = j if the i^{th} individual is censored at T_j the log likelihood contribution to the total likelihood is

$$\sum_{k<e_i} \log(1 - h_k) \tag{2.4}$$

Then the total log likelihood is given by

$$l = \sum_{deathi} \log h_{b_i} + \sum_{deathi} \left[\sum_{k<b_i} \log(1 - h_k) \right] + \sum_{censori} \left[\sum_{k<e_i} \log(1 - h_k) \right] \tag{2.5}$$

$$l = \sum_j d_j \log h_j + \sum_k \left[\sum_{j>k} d_j \right] \log\ 1 - h_k\ + \sum_k \left[\sum_{j \ge k} c_j \right] \log\ 1 - h_k \tag{2.6}$$

$$l = \sum_j \left[d_j \log h_j + \ n_j - d_j\ \log\ 1 - h_j\ \right] \tag{2.7}$$

where d_j is the number of observed death at τ_j, c_j is the number of censored at (τ_j , T_{j+1}); and n_j is the number of alive and uncensored at T_j. If h_j is the solution of

$$\frac{\partial l}{\partial h_j} = \frac{d_j}{h_j} - \frac{n_j - d_j}{1 - h_j} = 0 \qquad \text{then} \tag{2.8}$$

$$\hat{h}_j = \frac{d_j}{n_j} \tag{2.9}$$

This maximizes the likelihood since the total log likelihood function is concave down, so the K-M estimator of the survival function is

$$\hat{S}(t) = \prod_{\tau_j < t} \left(1 - \hat{h}_j\right) \tag{2. 10}$$

$$= \prod_{\tau_j < t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.11}$$

Therefore the K-M estimator is the maximum likelihood estimator. The K-M estimator deals with a discrete distribution. If the observations modeled come from unknown continuous distribution, the maximum likelihood estimator does not exist (Johansen, 1978).

Kaplan and Meier gave an approximate variance formular for their estimator which they attributed to Greenwood estimator (Greenwood, 1926). The Greenwood variance estimator for the K-M survival function is

$$\hat{V}ar_{Gr}\left[\hat{S}_{km}\right] = \hat{S}^2_{km}\left(\sum_{i=1}^{n} \frac{d_i}{n_i(n_i - d_i)}\right) \tag{2. 12}$$

The problem with Greenwood estimator is that it underestimates the true variance, especially in the right tails of the survival distributions (Peto, *et al* 1977). Furthermore it conveys little intuition about how the variance of the survival function changes, especially when censoring occur. Greenwood formular is only accurate for large sample, but concern have been noted about its little performance. Also the method tends to underestimate the standard error, especially in the right tails of the survival curve. To overcome this shortcoming Peto, *et al* (1977) introduced the estimator:

$$\hat{V}ar_p\left[\hat{S}_{km}\right] = \frac{\hat{S}^2_{km}\left[1-\hat{S}_{km}\right]}{n_i} \qquad (2.13)$$

But it is also observed that Peto‘s estimator underestimates the true variance, especially in the left and right tails of the survival distribution (Altman, 1977) and with heavy censoring.

To overcome the problem of censoring in variance, Zhao, (1996) gave the idea of homogenous variance estimator and Bork lop, (2005) adjusted hybrid variance estimator.

2.3 Comparison of Survival Distribution

The two survival curves can be compared statistically by testing the null hypothesis i.e. there is no significant difference in survival among the two or more groups. Example, survival curves of the patients receiving different treatments among males and females smokers. Comparison of two or more survival distribution (or curves) are common practice in medical research. Several methods are used for comparing survival distribution. The most common one used rank based tests called “log- rank test”. The test is also called

"Mantel – Cox test (Mantel, 1966). In rare cases some researchers may be interested in giving more weight to the time point where more number of subjects is available, then log-rank test is not appropriate. Other methods such as generalized Wilcoxon rank - sum test is used in this case (Hosmer and Lemeshaw, 1998). Wilcoxon rank-sum test is also known as Gehan test (Gehan, 1965).

Log rank method works on the same principle as Kaplan – Meier and thus, requires that survival duration is exactly available for both groups. Expected death at each time point in either group are obtained by following a procedure similar to the one followed for the contingency table. Total expected death for group I and group II are compared with the total observed deaths in group I and group II respectively and chi-square value with one degree of freedom is obtained. This is used to reject or fail to reject the null hypothesis of the equality of the survival curves.

The log rank test statistic consists of observed and expected events as χ^2 - test. To explain this, consider the two groups I and II and let $t_{(1)} < t_{(2)} <,...,< t_{(r)}$ be the r^{th} distinct death times for each group. At time t_j; let d_{1j} and d_{2j} be the number of deaths in group I and II respectively, and n_{1j} and n_{2j} are number of persons at risk prior to time t_j.

The log rank statistic is given by

$$\text{Log rank Statistic} = \frac{\left[\sum_{j=1}^{r} d_{1j} - EX_{1j}\right]}{Var} \sim \chi^2 \qquad (2.14)$$

where

$$E\ X_{1j} = \frac{n_{1j} d_j}{n_j} \qquad (2.15)$$

the mean of hyper geometric random variable and the variance of d_{1j} is:

$$Var(d_{1j}) = \frac{n_{1j} n_{2j} d_1 (n_j - d_j)}{n_j{}^2 (n_j - 1)}, \text{ therefore} \tag{2.16}$$

$$Var = \sum_{j=1}^{r} Var(d_{1j}) \tag{2.17}$$

When a large value of χ^2 is obtained, it will lead to the conclusion that two treatments are not equally effective. Gill, (1980) studied some properties of log rank test. The test is more powerful, reliable and appropriate as compared to other tests in a situation where two or more survival curves do not cross i.e. whose hazard functions are proportional. There are several alternative to log-rank test to test the equality of survival curves.

For example, the Gehan's Wilcoxon test is defined as:

$$\chi_G{}^2 = \frac{\left[\sum_{j=1}^{r} n_j \;\; d_{1j} - EX_{1j} \right]^2}{\sum_{j=1}^{r} n^2{}_j Var(d_{1j})} \sim \chi^2 \tag{2.18}$$

where n_j are weights whose values depends on the specific test.

The Wilcoxon test uses weights equal to risk size at t_j, $n_j = r_j$. This gives less weight to the longest survival times. Early failures receive more weight than later failures. The Wilcoxon test gives more emphasis on the information at the beginning of the survival curve where the number at risk is large. This type of weighing may be used to assess whether the effect of treatment on survival is the strongest in the earlier phases of administration and tend to be less effective over time. The log-rank test uses weight equal to one at $t_{(j)j}$ $n_j = 1$. This gives the same weight to each survival time. Therefore Wilcoxon

statistic is less sensitive than the log-rank statistic to difference of d_{1j} from E $[X_{1j}]$ in the tail of the distribution of survival times.

2.4 Semi-Parametric Approach

The Log-rank test is used to test whether there is statistical difference in survival times of the two or more groups, but the test cannot be used to test the effect of other independent or explanatory variables. Cox proportional hazard as semi-parametric model enable us to test the effect of other explanatory variables on survival times of different groups of patients, just like multiple regression model. Hazard is the dependent or response variable which is defined as probability of dying at a given time, assuming that the patients have survived up to that given time. Hazard ratio is an important term and is defined as the ratio of the risk of hazard occurring at any given time in one group compared with another group at that very time.

Cox proportional hazard model known as Cox PH model is a statistical technique for exploring the relationship between the survival of a patient and the explanatory variables. The model provides an estimate of the treatment effects on survival after adjustments for other explanatory variables. It allows estimate of hazards or risks to be made for an individual given their prognostic variables. Cox regression (Cox, 1972) is an indispensable tool in the field of survival analysis. It allows tests for differences in survival times of two or more groups of interest, while allowing adjustment for explanatory variables or covariates of interest. Cox regression model as semi parametric make fewer assumptions than typical parametric methods, but more assumptions than non parametric methods described earlier. In particular and in contrast, with parametric models, it makes no

assumptions about the shape of the base line hazard function. Cox regression model provides useful and easy way to interpret information regarding the relationship of the hazard function to predictors.

In diagnosis of a disease, medical doctors investigate the cause or other characteristics of a disease. For instance, the heart of a patient has the disease of high blood pressure, or family history of diabetes related to the development of diabetes disease. In this case, high blood pressure and family history of diabetes are referred to as explanatory variables or covariates or risk factors. Now the identification of the most important risk factors is becoming the important task for handling the disease. In regression analysis, generally we identify the risk factors. The presence of censoring in survival data will not allow us to use ordinary regression models. For this reason, Cox Proportional hazard or Cox regression model is widely used. Mohamed, *et al* (2007) compared Cox regression and parametric models (Weibull, Exponential, Log-normal) for survival of patients with gastric carcinoma, the results of their research showed that Cox model performed better than the other models. It is by far the most popular model for survival data analysis and it is accessible in a large number of statistical software packages (Altman and de Stavola, 1994).

The Cox Proportional Hazard model is given by

$$h(t/x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) \qquad (2.19)$$

where

$h_0(t)$ = baseline hazard function which is the hazard function for individual for whom all variables included in the model are zero.

x = $x_1, x_2, ..., x_p$ is the value of the vector explanatory variables for a particular individual.

$\beta^{/} = \beta_1, \beta_2, ..., \beta_p$ is a vector of regression coefficients, it describe the importance of covariates.

$\exp(\beta^{/} x)$ = The level of each individual's hazard

The model is based on the assumption that independent covariates affect the hazard in multiplicative way (Marubini and Valsecchi, 1995). X is a vector of covariates of interest. X may be discrete (sex, marital status etc), continuous (blood pressure etc) or the mixture of discrete and continous factors (interaction between height and sex). It is declared that the hazards for the two individuals (x_1 and x_2) are assumed to be proportionate. When comparing two groups, the hazard functions $h_1(t)$ and $h_2(t)$ can be determined for the groups. The hazard ratio is the quotient of the two hazard functions. Hazard ratio is a measure of how high the risk of an event is in group 1 in comparison to group 2. Group 2 is therefore considered to be the reference group. The hazard ratio is given by:

$$Hazard \quad Ratio \quad (HR) = \frac{h_1(t)\exp(\beta^{/} x_1)}{h_2(t)\exp(\beta^{/} x_2)} = \exp \ \beta_1(x_1 - x_2) \qquad (2.20)$$

If the hazard ratio > 1, this implies that the risk of an event in group 1 is higher than the risk of an event in group 2.

If the hazard ratio < 1, this implies that the risk of an event in group 1 is lower than the risk of an event in group 2.

If the hazard ratio = 1, then the risk of an event is approximately equal in both groups.

Cox assumes that hazard ratio (HR) comparing two specifications of covariates is constant over time. In fitting the Cox regression model, $h_0(t)$ and β are estimated. The approach is to attempt to maximize the likelihood function for the observed data

simultaneously with respect to $h_0(t)$ and β. A popular approach is proposed by Cox, (1975) in which a partial likelihood function that does not depend on $h_0(t)$ is obtained for β.

Partial likelihood is a method developed to make inference about regression parameters. Partial likelihood function can be constructed based on the proportional hazard model. The partial likelihood method for the Cox PH model which is a function of observed survival times and unknown parameters is given by:

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta' x_i(t_j))}{\sum_{k \varepsilon R(t_j)} \exp(\beta' x_k(t_j))} \tag{2.21}$$

where $x_i(t_j)$ be the vector of covariate is value for individual i who dies at time t_j.

It is should be noted that this likelihood function is only for uncensored individuals.

Let $t_1, t_2, ..., t_n$ be the observed survival time for n individuals and δ_i be the event indicator, which is zero if the i^{th} individual time is censored and unity otherwise. Now the likelihood function for the above equation can be expressed by:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' x_i(t_i))}{\sum_{k \varepsilon R(t_i)} \exp(\beta' x_k(t_i))} \right]^{\delta_i} \tag{2.22}$$

where $R(t_i)$ = risk set at time t_i.

The partial likelihood is valid when there are no ties in the data set. This means there is no two subjects who have the same event time.

Tied survival times can be handled by using the likelihood form of Breslow, (1974). This is simple to understand and is more suitable in a situation where the number of tied

observations at any death time is not too large. The model, dealing with tied situation is explained by Efron, (1977) and Kalbkleisch and Prentice (1980).

In case of large tied survival times, Cox's assumes the discrete survival times and generalized the model by introducing the concept of logistic transformation.

$$\frac{h(t)dt}{1-h(t)dt} = \frac{h_0(t)dt}{1-h_0(t)dt}\exp(\sum_{j=1}^{k}\beta_j x_{ji}) \qquad (2.23)$$

All these procedures are used for the Cox Proportional hazard model. If the assumptions of Cox PH model fail, one cannot use the procedures. The main assumption of Cox PH model is proportional hazard. Proportional hazard means that the hazard function of one individual is proportional to the hazard function of the other individual, that is, the hazard ratio is constant over time. The most common reason for the failure of the Cox PH model assumption is the involvement of the time dependent covariates. Some of the examples of the dependent covariates are, status of a woman may change during the period of study i.e. girl become married woman and married woman become mother, regular examination of patients in a clinic etc. In time dependent Cox regression, the values of variables changes over the period.

Time dependent covariates are also known as updated covariates (Altman and de Stavola, 1994). There are two types of time dependent covariates. These are external variable and internal variable, according to Kalbfleisch and Prentice, (1980).

External variable is the one in which its value in most cases is known in advance. Such variable does not require the survival of the patients. Examples of external variable

are age of patient, dose of a drug etc. Internal variable measurement is only possible if the patient is alive. Examples are white blood counts, systolic blood pressure.etc.

2.5 Parametric Approach

The parametric regression models used in this study are Weibull regression and Gompertz regression models. These models are among the major models that have been used for modeling survival or failure times.

2.5.1 Weibull Regression Model

Weibull model plays the same role in the analysis of survival data as in the case of normal distribution in linear modeling. Weibull model is more general and flexible than exponential model and allow for hazard rates that are non constant but monotonic.

As a result of versatility in fitting time to failure of a very extensive variety to complex mechanisms, the Weibull distribution has lately assumed the centre stage especially in the field of life testing and reliability analysis. Weibull distribution has wide variety of applications in many areas. The most used methods which are considered to be the traditional methods are maximum likelihood and the moment estimation (Cohen and Whitten, 1982). Bijan, *et al* (2008), investigated the comparative performance of semi parametric Cox model and parametric (Weibull, Exponential, Log-normal) models in survival analysis of patients with stomach carcinoma and found Weibull model as the best model in multivariate analysis. Hui, (2011) compared Weibull and Cox models in a study of patients with gastric cancer and found Weibull model, best in fitting the data compared to Cox model.

The model is characterized by positive parameters (Collett, 1994), the parameter λ is known as a scale parameter and P the shape parameter, because it determines whether the hazard is increasing, decreasing or constant over time. In Weibull model, if the intercept and slope are roughly estimated as log (λ) and P and the lines are parallel, then the proportional hazard model is valid. The Weibull tends to exponential distribution if the shape parameter P equals to1.

The shape parameter of the Weibull model works in the following ways:

If P < 1, then the hazard is monotonically decreasing with time.

If P > 1, the hazard is monotonically increasing with time.

If P = 1, the hazard is flat i.e constant hazard rates, and the model is exponentiated, i.e. Weibull model nests the exponential model. This means Weibull model can be used as a test to see if the exponential model is appropriate.

The survival function of the Weibull model is $S(t) = \exp(-\lambda t)^p$ (2.24)

The hazard function is: $h(t) = \lambda p(\lambda t)^{p-1}$, (2.25)

where $\lambda = \exp(x_i \beta)$ (2.26)

The probability density function is

$$f(t) = \lambda p(\lambda t)^{p-1} \exp(-\lambda t)^p \quad (2.27)$$

The expected duration from Weibull model is $E(T) = \left(\frac{1}{\lambda}\right)^{\frac{1}{p}} \Gamma\left(1+\frac{1}{p}\right)$ (2.28)

where Γ denotes gamma function

The variance is $Var(T) = \left(\frac{1}{\lambda}\right)^{\frac{2}{p}} \left\{\Gamma\left(1+\frac{2}{p}\right) - \Gamma\left(1+\frac{1}{p}\right)^2\right\}$ (2.29)

2.5.2 Gompertz Regression Model

The Gompertz distribution model plays an important role in modeling human mortality and fitting actuarial tables. Gompertz distribution is one of the most important growth models. Gompertz model was formulated by the British actuary Benjamin Gompertz in 1825; the model was used frequently by medical researchers and biologists in modeling mortality ratio data. According to the Gompertz (1825) law, the number of individuals alive in a human population decreases exponentially with age at an exponentially increasing rate. As Gompertz and many others have pointed out, that model serves well only for persons between about 25 and 85 years of age.

The discrepancy outside that age is not evident in the ordinary survival plot but is obvious in plots of mortality rate against age. In their survey, Olshansky and Carnes (1997), considered extensively 'non Gompertzian mortality at older ages.' In their reports, the older members of the population depart erratically from the classical model, but overall survival is better than predicted by the standard Gompertz law.

Gompertz is a growth model and has been used in relation with tumour development. Ahuja and Nash (1967), showed that Gompertz distribution was with a simple conversion, related to some distribution in the Pearson distribution family. Gompertz distribution is a continuous probability distribution and it is often applied to describe the distribution of adults' lifespan by the demographers and actuaries. Elvan, *et al* (2010) conducted a research to compare five models (Gompertz, Gamma, Log-Logistic, Weibull and Log-normal) using breast cancer data. The results of the study showed that Gompertz model was the most suitable model to fit their data. Anet and Nestor, (2011) conducted a research on the application of hazard models for patients with breast cancer in

Cuba using 6381 patients to compare six models (Gompertz, Exponential, Weibull, Log-logistic, Lognormal, and Generalized Gamma). Their results showed that Gompertz model fit the data well.

The analysis of growth data becomes more important in many fields of study. Biologists are interested in the description of biological growth and in trying to understand its underlying biological process. Chemists are interested in the formulation of the product of a reaction of two or more chemical substances. In agriculture there are obvious economic and management advantages in knowing how large things grow, how fast they grow and how does these factors respond to environmental conditions or treatment over time. Social scientists are interested, for example, in the growth of populations, new political parties, the foods supply and energy demand. The same type of model occur when the explanatory variable x is no longer time but increasing intensity of some other factors, such as the growth of smog (mixture of water vapour and air) with increase in solar radiation, weight gains with increased nutrient in diet, changes in crop yield with increase density of planting and so on.

Related fields of science such as population studies, animal growth, and gerontology also considered the Gompertz distribution for the analysis of survival. The model was used as a growth model for organism by Wright, (1986), and its properties were formally developed by Winsor, (1932). Its initial formulation was largely empirical, but later Medawar, (1940) derived it as a growth model for the heart of a chicken. In veterinary literature the Gompertz function is frequently used as a growth model for birds (Aggrey, 2002).

Computer scientists model the failure rates of computer codes by the Gompertz distribution. In marketing science, it has been used as an individual level model of customer lifetime. Jaheen, (2003), used maximum likelihood estimations method to obtained Gompertz distribution parameters λ and P, in order to derive the composite Gompertz distribution which is assumed to accord with the Gamma distribution with one of the two parameters being random. Gompertz is useful for monotone hazard rates that are exponentially increasing or decreasing with time. The shape parameter works in the following ways:

If P < 1, then the hazard is monotonically decreasing with time

If P > 1, then the hazard is monotonically increasing with time

If P = 1, then the hazard is flat and the model is exponentiated.

The survival function of the Gompertz model is:

$$S(t) = \exp\left[\frac{\lambda}{p} \; 1 - \exp(pt) \; \right] \tag{2.30}$$

The hazard function is $h(t) = \lambda \exp(pt)$ (2.31)

The probability density function is:

$$f(t) = \lambda \exp(pt) \exp\left(\frac{\lambda}{p} (1 - \exp(pt)) \right) \tag{2.32}$$

The expected duration from Gompertz model is $E(T) = \frac{1}{\lambda} G^{\frac{\lambda}{p}}$ (2.33)

where $G = \int_x^{\infty} \frac{1}{y} \exp(-y) dy$ (2.34)

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The data on 312 cases of breast cancer, used in this study were obtained from Ahmadu Bello University Teaching Hospital, Shika, Zaria from the period January 1997 to December 2012.

The different probabilistic survival analysis methods which include parametric, semi-parametric and non parametric models used in the analysis of the data are:

-Non parametric – Kaplan-Meier or Product Limit Estimator

-Semi-Parametric – Cox Proportional Hazard Model

-Parametric: - Weibull and Gompertz Regression Models.

The above models are to be considered because they are widely used in fitting survival data. The models are to be fitted to the data with a view to find the best fit.

The statistical package "Stata "version 12 was used for analysing the data.

3.2 Kaplan-Meier Estimator

In cancer trial, Kaplan-Meier (K-M) method is the recommended technique in survival analysis. K-M is the most popular in developing survival function (Collectt, 2003). The method is used to measure the fraction of subjects living for a certain period of time after treatment. It is applied by analyzing the distribution of patients' survival times following their recruitment to a study. The analysis expresses these in terms of proportion of patients still alive up to a given time, following their recruitment. In terms of graph, a plot of proportion of patients' surviving against time has a characteristic decline; the steepness of the curve indicates the efficacy (or ability to produce a desired result) from the

treatments being investigated. The shallower part of the curve shows the more effective treatment. In analysing the survival data, two functions that are dependent on time are of particular interest; the survival function and the hazard function. The survival function denoted by

S(t) is the probability of surviving at least to time t. The hazard function denoted by h(t) is the conditional probability of dying at time t having survived to that time. The graph of

S(t) against t is called the survival curve.

The Kaplan-Meier method can be used to estimate this curve from the observed survival times without the assumption of the underlying probability distribution. The method is based on the basic idea that the probability of surviving P or more periods from entering the study is the product of the P observed survival rates for each period i.e. the cumulative surviving; and is given by:

$$S(p) = (K_1)\,(K_2)\,(K_3),\ldots,(K_p) \qquad (3.1)$$

where

K_1 = Proportion of surviving the first period

K_2 = Proportion of surviving beyond the second period conditional on having survived up to the second period and so on.

The proportional surviving period i having survived up to period i is given by

$$K = \frac{r_i - d_i}{r_i} \qquad (3.2)$$

where

r_i = the number alive at the beginning of the period

d_i = The number of deaths within the period

3.3 Comparison of the Survival Curves of Two Groups Using the Log-rank Test

A statistical hypothesis test called Log-rank test is used to compare the two survival curves. It is used to test the null hypothesis that there is no difference between the survival curves, i.e. the probability of event occurring at any point in time is the same for each population.

The test statistic is given as

$$\text{Log-rank test statistic} = \frac{(0_1 - E_1)^2}{E_1} + \frac{(0_2 - E_2)^2}{E_2} \tag{3.3}$$

where

0_1 and 0_2 are total number of observed events in group 1 and group 2 respectively and E_1 and E_2 are total number of the expected events in group 1 and group 2 respectively.

The total expected number of events for a group is the sum of the expected number of events at the time of each event. The expected number of events at the time of an event can be calculated as the risk for death at that time multiplied by the number alive in the group. Under the null hypothesis, the risk of death, i.e. number of deaths divide by the number alive can be calculated from the combined data for those groups.

$$E_2 = \sum_{i=1}^{k} \frac{d_i}{r_i} r_{2i} \tag{3.4}$$

where r_{2i} = the number alive from group 2 at the time of event i

E_1 is calculated as $n - E_2$, where n = the total number of events

The test statistic is compared with a χ^2 - distribution with 1 degree of freedom.

3.4 Cox Proportional Hazard Model

Cox Proportional Hazard Model or Cox regression model abbreviated as Cox (PH) is a well recognized statistical technique for exploring the relationship between the survival of a patient and the several explanatory variables or covariates. Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. It allows an estimate to be made on the hazard (or risk) of death, or other event of interest, for individuals given their prognostic (or forecast of a disease) variables.

The principle of Cox Proportional hazard model is to link the survival time of an individual to covariates. For instance, in medical activity, we are interested in finding out which covariate has the most important impact on the survival time of a patient. Interpreting a Cox model involves examining the coefficient for each explanatory variable. A positive regression coefficient for an explanatory variable means that the hazard for patient having a high positive value on that particular variable is high. On the other hand, a negative regression coefficient implies a better prognostic for patients with higher value of that variable. Cox method does not assume any particular distribution for survival times, rather it assumes that the effect of different explanatory variables on survival are constant over time and are additive in a particular way. The model assumes that the hazard for two individuals say x_1 and x_2 is proportional. In other words, the hazard of an individual is proportional to the hazard of any other individual.

The hazard function is the probability that an individual will experience an event, for example death with a small time interval, given the individual has survived up to the beginning of the interval. It can be interpreted as the risk of dying at time t.

The hazard is modeled as:

$$h(t/x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) \quad (3.5)$$

where

$x_1,\ldots, x_p$ = collection of explanatory variables.

$h_0(t)$ = baseline hazard time t representing the hazard for a person with value 0 for all explanatory variables.

$\beta_1,...,\beta_p$ = regression coefficients describing the importance of the covariates, which is estimated by the partial likelihood estimation procedure.

3.5 Cox Proportional Hazard Model Partial Likelihood Estimation

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model, when applied to a data set and given a statistical model; maximum likelihood estimation provides estimates for the model's parameters.

Parameter estimates in Cox model are obtained by maximizing the partial likelihood function for the observed data simultaneously with respect to $h_0(t)$ and β. A popular approach is proposed by Cox, (1975), he proposed that the partial likelihood does not depend on $h_0(t)$ which is obtained for β. Partial likelihood is a procedure developed to take

inferences about the regression parameters in the presence of nuisance parameters i.e. $h_0(t)$ in Cox model.

To estimate the model, let $t_1, t_2, ..., t_n$ be the observed survival time for n individuals, let the ordered death time r individual be $t_1 < t_2 <, ..., < t_r$ and let R_{t_j} be the risk set just before

t_j and r_j for its size, so that R_{t_j} is the group individuals alive and uncensored at time just prior to t_j.

The conditional probability that individual dies at t_j given that one individual from the risk set R_{t_j} dies at t_j is (Jiezhi, 2009)

P (individual i dies at t_j /one death from the risk set R_{t_j} at t_j)

$$= \frac{P(individual\ i\ dies\ at\ t_j)}{P(one\ death\ at\ t_j)}$$

$$= \frac{P(individual\ i\ dies\ at\ t_j)}{\sum_{k \in R_{t_j}} P(individual\ k\ dies\ at\ t_j)}$$

$$= \frac{\exp(\beta' x_i(t_j))}{\sum_{k \in R_{t_j}} \exp(\beta' x_i(t_j))} \qquad (3.6)$$

Now, the partial likelihood function of Cox model is given as

$$L(\beta) = \prod_{j=1}^{r} \left[\frac{\exp(\beta' x_i(t_j))}{\sum_{k \in R_{t_j}} \exp(\beta' x_i(t_j))} \right] \qquad (3.7)$$

In which $x_i t_j$ is the vector covariate values for individual i who dies at t_j.

The log partial likelihood is given as:

$$LogL(\beta) = \sum \left\{ \beta x_i - \log\left(\sum_{j \in R_{ij}} \exp(\beta' x_i) \right) \right\} \tag{3.8}$$

The partial likelihood is valid when there are no ties in the data set, but often ties occur in continuous survival data that are collected in days, weeks and months. If there are only few ties Breslow, (1974) gave approximation as:

$$L(\beta) = \prod_{j=1}^{r} \left\{ \frac{\prod_{j \in Dt_j} \exp(\beta' x_i)}{\left[\sum_{j \in R_{t_j}} \exp(\beta' x_i) \right]^{d_i}} \right\} \tag{3.9}$$

where

D_{tj} = set of individual failing at t_j

d_i = number of failure occurring at t_j

3.6 Weibull Regression Model

Weibull model is suitable for analysing survival data. It has been recognized as an appropriate model in reliability studies and life testing problems such as time to failure or life length of a component or product. The Weibull distribution is characterized by two parameters, one is the shape parameter P (dimensionless) and the other is the scale parameter λ.

Weibull as parametric model assume that the survival time follow a specified probability distribution and has an explicit relationship with covariates. i.e assumption is made on the shape of the underlying hazard function (Lee and Wang, 2003). Parametric methods of survival analysis assumed a distribution of hazard rates as a function of time. The methods also assume that censoring observations should be independent.

Over the years, estimation of the shape and scale parameters for Weibull distribution function has been approached through maximum likelihood method, which is considered to be the traditional method of estimation (Cohen and Whitten, 1982). Hossain and Zimmer, (2003) obtained the maximum likelihood estimate of Weibull parameters for complete and censored samples.

Let $t_1, t_2, ..., t_n$ be a random sample from Weibull distribution with probability density function (pdf).

$$f(t/\lambda, p) = \left(\frac{p}{\lambda}\right)\left(\frac{t}{\lambda}\right)^{p-1} \exp\left(\frac{-t}{\lambda}\right)^{p} \tag{3.10}$$

Taking the likelihood of the function (Saunders, 2007)

$$L(t/\lambda, p) = \prod_{i=1}^{n} \left(\frac{p}{\lambda}\right)\left(\frac{t_i}{\lambda}\right)^{p-1} \exp\left(\frac{-t_i}{\lambda}\right)^{p} \tag{3.11}$$

λ and p are the parameter to be estimated, then the natural logarithm of the likelihood function is:

$$\ln L(t/\lambda, p) = n\ln(p) - np\ln(\lambda) + (p-1)\sum_{i=1}^{n}\ln(t_i) - \sum_{i=1}^{n}\left(\frac{t_i}{\lambda}\right)^{p} \tag{3.12}$$

Differentiating with respect to p and λ in turn and equate the expression to zero

$$\frac{\partial \ln(t/\lambda, p)}{\partial p} = \frac{n}{p} + \sum_{i=1}^{n}\ln(t_i) - \frac{1}{\lambda}\sum_{i=1}^{n} t_i^{\,p} \ln p = 0 \tag{3.13}$$

$$= \frac{-n}{p} + \frac{1}{\lambda^2}\sum_{i=1}^{n} t_i^{\,p} = 0 \tag{3.14}$$

In eliminating λ between (3.14) and (3.15), we simplify to get

$$\frac{\sum_{i=1}^{n} t_i^{\,p} \ln t_i}{\sum_{i=1}^{n} t_i^{\,p}} - \frac{1}{P} - \frac{1}{n}\sum_{i=1}^{n} \ln t_i = 0 \tag{3.15}$$

This can be solved to get estimate of p using standard iterative procedure of Newton-Raphson method.

$$P = \left[\frac{\sum_{i=1}^{n} t_i^{\,p} \ln t_i}{\sum_{i=1}^{n} t_i^{\,p}} \frac{1}{n}\sum_{i=1}^{n} \ln t_i \right]^{-1} \tag{3.16}$$

Once P is determined, λ can be estimated using

$$\frac{-n}{p} + \frac{1}{\lambda^2}\sum_{i=1}^{n} t_i^{\,p} = 0 \tag{3.17}$$

and

$$\lambda = \left[\frac{1}{n}\sum_{i=1}^{n} t_i^{\,k} \right]^{\frac{1}{p}} \tag{3.18}$$

3.7 Gompertz Regression Model

The Gompertz distribution has many applications, particularly in medical and actuarial studies. Gompertz as a parametric model also assumed that the survival times follow particular distribution. Another basic assumption of Gompertz law is that the resistance to death decline exponentially with age and the model assumes independent of censoring. Related field in such as gerontology considered Gompertz distribution for analysis of survival (Economos 1982).

Maximum-likelihood estimation (MLE) is a method useful for fitting mortality parameters. This approach, defines for each possible value of the Gompertz parameter the likelihood of observing the data i.e the particular times of death, which occurred, if the mortality rates were given exactly by a Gompertz curve with those parameters λ and P. For $0 \leq t < \infty$ and $\lambda > 0$, the parameter P determine the shape of the hazard function. When P = 0, the survival time become an exponential distribution, i.e the exponential distribution is also a special case of the Gompertz distribution.

The probability density function (Pdf) of Gompertz distribution is given by:

$$f(t) = \lambda \exp(pt) \exp\left(\frac{\lambda}{p}(1 - \exp(pt))\right) \tag{3.19}$$

then

Taking the likelihood of the function

$$Lf(t/\lambda, p) = \prod_{i=1}^{n} \lambda \exp(p\sum_{i=1}^{n} ti) \exp\left(\frac{\lambda}{p}(1 - \exp(p\sum_{i=1}^{n} ti))\right) \tag{3.20}$$

Taking the natural logarithm of the function gives

$$\ln L(t/\lambda, P) = \ln \lambda + P\sum_{i=1}^{n} t_i + \frac{\lambda}{P}(n - \exp(P\sum_{i=1}^{n} t_i)) \tag{3.21}$$

Differentiating with respect to p and λ in turn and equating the expressions to zero

$$\frac{\partial \ln L(t/\lambda, P)}{\partial \lambda} = \frac{1}{\lambda} + \frac{1}{P}(n - \exp(P\sum_{i=1}^{n} t_i)) \tag{3.22}$$

$$\frac{1}{\lambda} + \frac{(n - \exp(P\sum_{i=n}^{n} t_i))}{P} = 0 \tag{3.23}$$

$$\frac{P + \lambda(n - \exp(P\sum_{i=n}^{n} t_i))}{P\lambda} = 0 \tag{3.24}$$

$$P + \lambda(n - \exp(P\sum_{i=n}^{n} t_i)) = 0 \tag{3.25}$$

hence,

$$\lambda = \frac{-P}{n - \exp(P\sum_{i=1}^{n} t_i)} \tag{3.26}$$

Similarly for P,

$$\frac{\partial \ln L(t/\lambda, P)}{\partial P} = \sum_{i=1}^{n} t_i - \lambda \sum_{i=1}^{n} t_i \exp(P\sum_{i=1}^{n} t_i) - \frac{n\lambda}{P^2} + \frac{\lambda}{P^2}\exp(P\sum_{i=1}^{n} t_i) \tag{3.27}$$

$$P^2\sum_{i=1}^{n} t_i - \lambda P^2 \sum_{i=1}^{n} t_i \exp(p\sum_{i=1}^{n} t_i) - n\lambda + \lambda \exp(p\sum_{i=1}^{n} t_i) = 0 \tag{3.28}$$

$$P^2\sum_{i=1}^{n} t_i(1 - \lambda \exp(p\sum_{i=1}^{n} t_i) = \lambda(n - \exp(p\sum_{i=1}^{n} t_i)) \tag{3.29}$$

$$P^2 = \frac{\lambda(n - \exp(p\sum_{i=1}^{n} t_i))}{\sum_{i=1}^{n} ti(1 - \lambda \exp(p\sum_{i=1}^{n} t_i))} \tag{3.30}$$

$$P = \left[\frac{\lambda(n - \exp(p\sum_{i=1}^{n} t_i))}{\sum_{i=1}^{n} t_i(1 - \lambda \exp(p\sum_{i=1}^{n} t_i))} \right]^{\frac{1}{2}} \quad (3.31)$$

3.8 Akaike Information Criterion (AIC)

In statistical modeling, one of the main challenges is to select a suitable model from a candidate family to characterize the underlying data. Model selection criteria provide a useful tool in this regard. Akaike Information Criterion (AIC) is a technique that measures the goodness of an estimated statistical model and selects a model from a set of candidate models. The chosen model is the one that is expected to minimize the difference between the model and the truth. Given a set of data, several competing models may be ranked according to their corresponding AIC, and the one having the smallest AIC is the best. AIC was the first model criterion to gain widespread acceptance. AIC was introduced in 1973 by Hirotogu Akaike as an extension to the maximum likelihood principle. AIC is given by the formular:

AIC = -2{ln (likelihood)} + 2k

where

Likelihood = the probability of the data in a given a model

K = the number of the parameter in the model

CHAPTER FOUR

ANALYSIS AND DISCUSSION OF RESULTS

4.1 Introduction

In this chapter, the results obtained from the analysis of the data were discussed. The data collected for the analysis are attached as appendix I. The K-M approach was used to describe the survival functions of the breast cancer patients and the Log-rank tests were used to compare the survival curves of the patients. The breast cancer data were used to fit the three models viz: Cox Proportional Hazard Model, Weibull Regression Model and Gompetz Regression Model. The models were evaluated using Akaike Information Criterion (AIC) to determine the best model among the three. The model with lower AIC value is usually considered a better model in terms of goodness of fit of the breast cancer data.

4.2 Kaplan –Meier (K-M) Survival Function

K-M was used to describe how to estimate and graph the survival curves of breast cancer patients at Ahmadu Bello University Teaching Hospital Zaria, between 1997 and 2012. The mean age of the breast cancer patients was found to be 43.39 years with the standard deviation of 11.74 years. From figure 4.1, the overall median survival time is 10

months. This indicates that 50% of the breast cancer patients survived less or equal to 10 months and the other 50% survived longer than 10 months after they are diagnosed with the disease. This is the survival time at which the cumulative survival function is equal to 0.5. The 5-year overall survival rate was found to be approximately 35.0%. This rate is lower than the survival rates of other countries like India, China and Malaysia with survival rates

of 54.2%, 49.5% and 45.1% respectively. (Nor, *et al* 2013). In Figure 4.2, the median survival time for the housewives is 16 months and for the other occupations is 10 months. These clearly shows that 50% of the breast cancer patients who are housewives survived longer than 16 months and 50% of the patients who are in other occupations survived longer than 10 months. In figure 4.3, the median survival time for age group 20-39 was found to be 19 months and for age groups 40-49, 50-59 and 60 years and above was 7, 5 and 10 months respectively. This indicates that 50% of the breast cancer patients age 20-39 survived longer than 19 months. 50% of age groups 40-49, 50-59 and 60 years and above survived longer than 7, 5 and 10 months respectively.

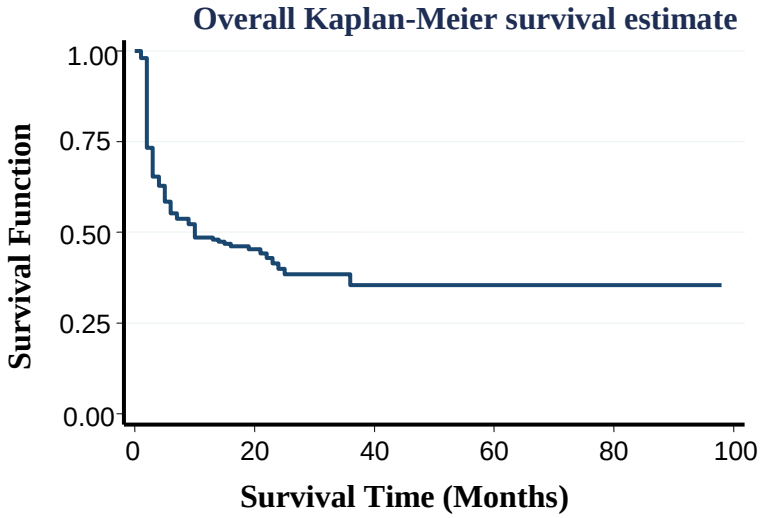


Figure 4.1 Kaplan-Meier (K-M) Curve for overall survival estimate

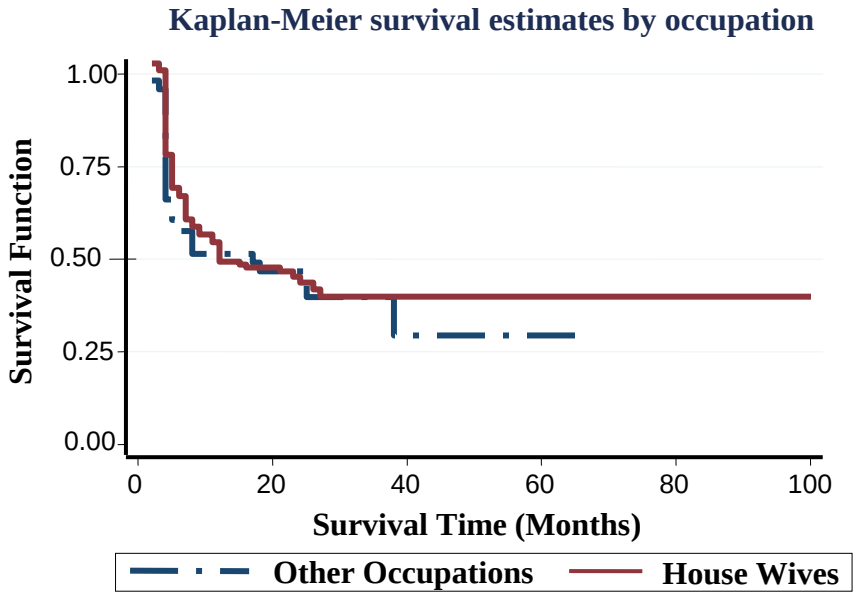


Figure 4.2 Kaplan – Meier (K-M) curves for survival estimate of occupation

Survival Function

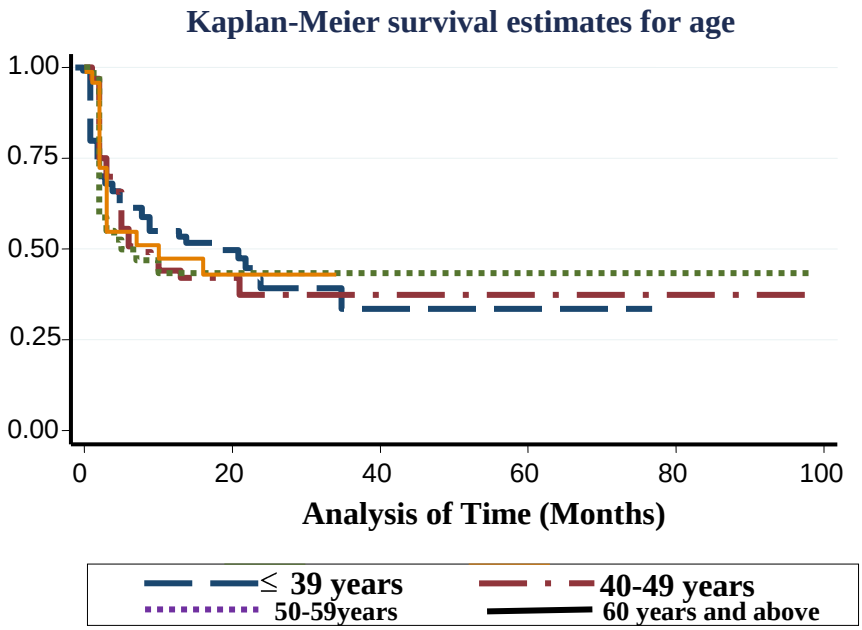


Figure 4.3 Kaplan-Meier (K-M) Curves for survival estimate of age

4.3 Log- rank Tests for Equality of Survival Functions (Survival Probabilities)

To describe how to evaluate whether or not K-M curves for two or more groups are statistically significant, a popular testing method called Log-rank test is used and the test used chi-square statistic. When two K-M curves are statistically the same or equivalent it means that based on testing procedure that compares the two curves in some overall sense, we do not have evidence to indicate that the true population survival curves (probabilities) are different. The Log-rank statistic, like any other statistics used in other kinds of chi-square tests, makes use of observed versus expected cells counts over categories of outcomes. The categories for Log-rank statistic are defined by each of the ordered failure times for the entire set of data being analysed.

Table 4.1 Stages of the Breast Cancer Patients

Stages	Events Observed	Events Expected
I	0	9.15
II	3	84.45

III	147	56.40
Total	150	150.00

Chi-square = 300.23 P-value = 0.0001

H_0: The breast cancer patients for the three stages have the same survival probabilities.

H_1: The breast cancer patients for the three stages do not have the same survival probabilities.

The log-rank statistic is 300.23 and the P-value is 0.0001, this indicates that the null hypothesis (H_0) can be rejected, and that the breast cancer patients in the three stages do not have the same survival probabilities. In other words, the survival experiences of the breast cancer patients are not the same with respect to their stages of the cancer at 1% level of significance.

Table 4.2 Occupations of the Breast Cancer Patients

Occupations	Events Observed	Events Expected
Housewives	109	108.94
Other occupations	41	41.06
Total	150	150.00

Chi-square = 0.01 P-value = 0.9899

$H_{0:}$ The breast cancer patients for housewives and other occupations have the same survival probabilities.

H_1: The breast cancer patients for housewives and other occupations do not have the same survival probabilities.

The log-rank statistic is 0.01 and the P-value is 0.9899, this indicates that the null hypothesis (H_0) cannot be rejected. This means that the survival probabilities of the breast cancer patients are almost the same for housewives and other occupations at 1% level of significance. House wives in this study refer to the patients who are not engaged in any work or business. For other occupations, the patients are engaged in some kind of work or business.

Table 4.3 Age groups of the Breast Cancer Patients

Age groups	Events Observed	Events Expected
≤ 39years	54	61.57
40-49 years	47	44.63
50-59 years	32	26.95
60 years and above	17	16.85
Total	150	150.00

Chi-square = 8.81 P-value = 0.0378

H_0: The breast cancer patients in four age groups have the same survival probabilities.

H_1: The breast cancer patients in four age groups do not have the same survival probabilities.

The log-rank statistic is 8.81 and the P-value is 0.0378, this indicates that the null hypothesis (H_0) cannot be rejected, and that the breast cancer patients in the four age groups have the same survival probabilities, the survival experience of the breast cancer patients are the same with respect to their age groups at 1% level of significant.

Table 4.4 Results of the Treatment of the Breast Cancer Patients

Results	Events Observed	Events Expected
Improved Condition	38	99.41
Unimproved Condition	112	50.59
Total	150	150.00

Chi-square = 139.11 P-value = 0.0001

H_0: The breast cancer patients have the same survival probabilities for the results of the treatment.

H_1: The breast cancer patients do not have the same survival probabilities for the results of the treatment

The log-rank statistic is 139.11 and the P-value is 0.0001, this indicates that the null hypothesis (H_0) can be rejected, and that the breast cancer patients do not have the same survival probabilities, the survival experience of the breast cancer patients are not the same with respect to the results of their treatment at 1% level of significant.

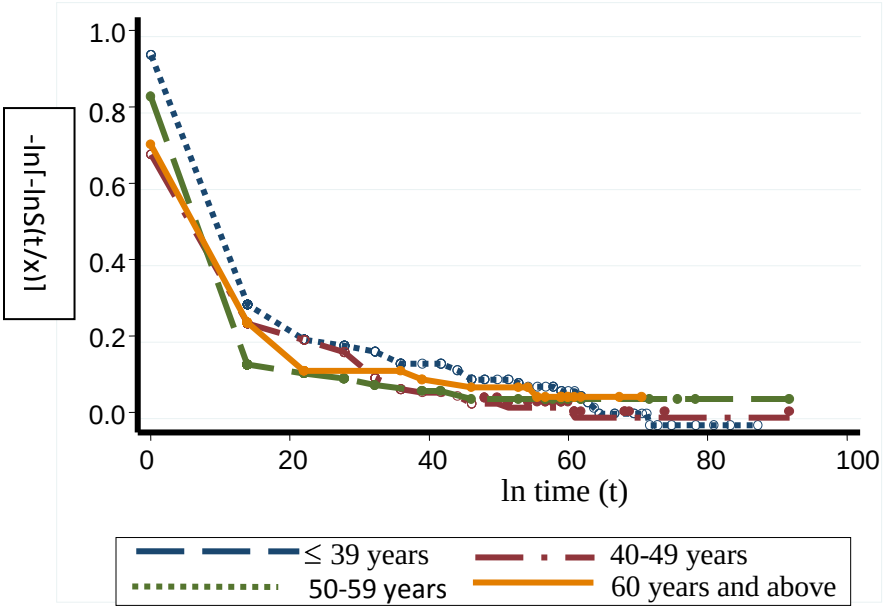
Table 4.5 Summary of Log-rank Tests for Group Variables

Group Variables	Tests Statistics	P-Values
Stages	300.23	0.0001
Occupations	0.01	0.9899
Age Groups	8.81	0.0378
Results of the Treatment	139.11	0.0001

4.4 Tests for Proportional Hazard Assumption

The Proportional assumption requires that the hazard ratio (HR) which is the measure of effects is constant over the time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual, the proportionality constant is independent of time. The general rule is that if the hazard curves are approximately parallel then proportional assumption is met, but when the hazard curves crossed in the graph, the proportional assumption is violated, so the Cox proportional model is not appropriate (Stephen, 2009).

Consider figure 4.4 for the test of proportional hazards assumption for age. Figure 4.5, 4.6 and 4.7 displayed the tests for proportional assumption for stages of the disease, results of the treatment and occupation of the breast cancer patients' respectively.



S(t/x) = Kaplan – Meier Estimate of Survival Function

Figure 4.4 Test of Proportional Hazard Assumption for Age Groups

From figure 4.4, it can be observed that the hazard curves in the graph crossed each other; the proportional assumption is violated for age group.

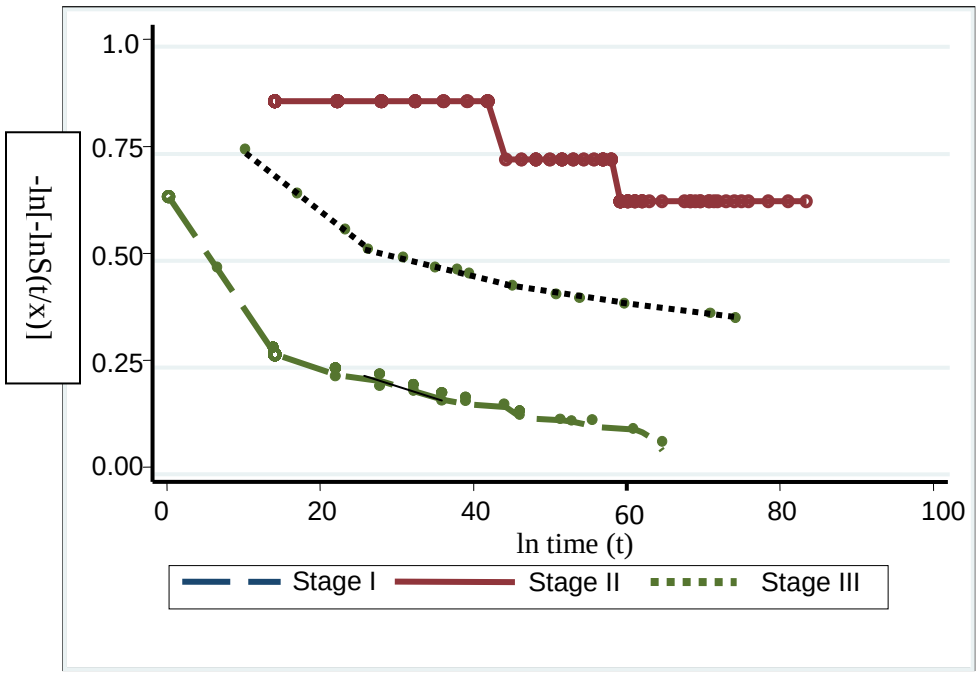


Figure 4.5 Test of Proportional Hazard Assumption for Stages of Breast Cancer

From figure 4.5, it can be observed that the hazard curves in the graph did not cross each other; the proportional assumption is met for stages of breast cancer.

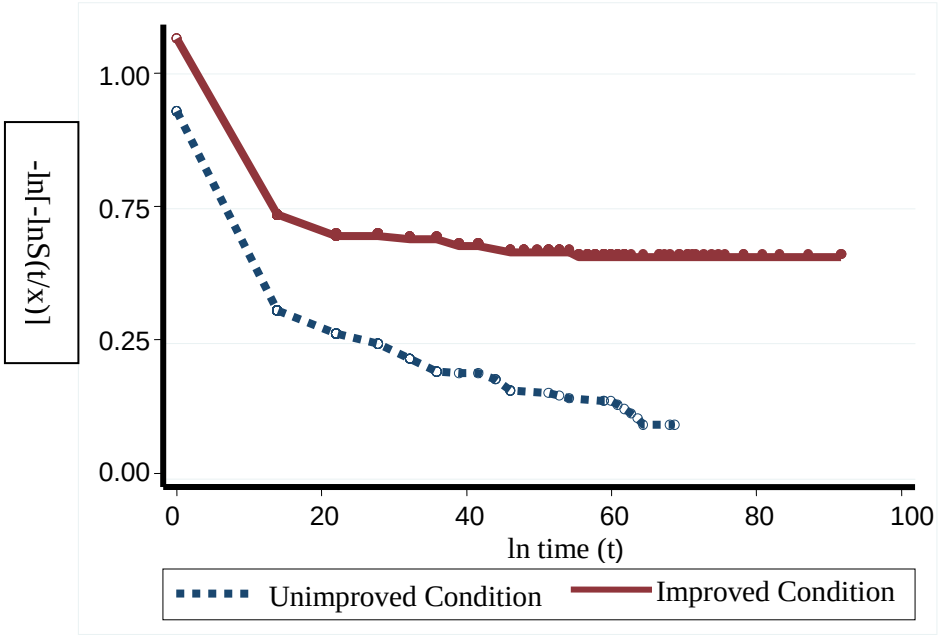


Figure 4.6 Test of Proportional Hazard Assumption for Results of the Treatment

From figure 4.6, it can be observed that the hazard curves in the graph did not cross each other; the proportional assumption is met for results of the treatment.

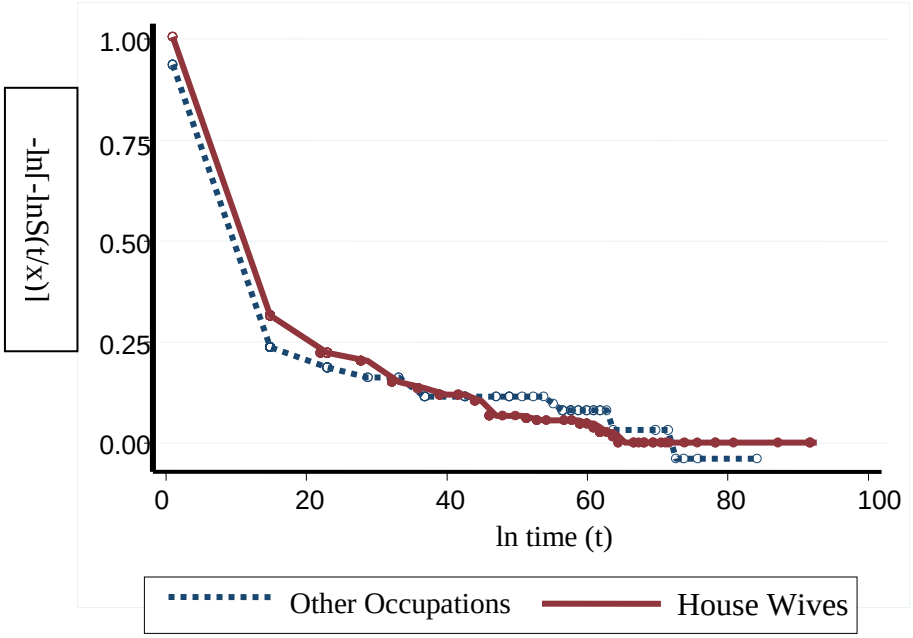


Figure 4.7 Test of Proportional Hazard Assumption for Occupation of the Patients

From figure 4.7, it can be observed that the hazards curves crossed each other and therefore the assumption is violated for occupation of the patients.

4.5 Results from the Analysis of Cox Proportional Hazard Model

Log likelihood = -2.64.32742

Table 4.6 Results for Cox Model

Covariates	Hazard Ratio (HR)	Std. Error	P-values	[95% Conf. Interval]	
Occupation (H/W)	0.8501	0.1623	0.0390	1.5847	2.2358
Treat.(improved)	0.7607	0.2155	0.0001	1.1779	1.8977
Stage II	1.1385	1.3423	0.1330	0.6785	2.0012
Stage III	2.8717	1.2165	0.0000	2.2709	3.4413
Age ≤ 39	0.6071	0.2228	0.0440	0.2957	0.5465
Age 40-49	0.4900	0.3291	0.0270	0.5969	1.8997
Age50-59	0.3465	0.2314	0.1740	0.1245	1.6755

The hazard ratio computed for occupation is 0.85, 95% CI: 1.58 – 2.24, P-value = 0.039. This indicates that housewives are 0.85 times less likely to be at the risk of death compared to patients in other occupations. But occupation of the patients does not have significant effect on the mortality of the patients at 1% level of significance. 1% level of significant is selected because breast cancer is a medical data; therefore the error margin has to be as small as 1%.

The hazard ratio for results of the treatment is 0.76, 95% CI: 1.18 – 1.89, P-value = 0.0001. This indicates that breast cancer patients with improved condition are 0.76 less likely to be at the risk of death compared to patients with unimproved condition. This means condition of the patient is found to have significant effect on the mortality of the patients at 1% level of significance.

The hazard ratio for stage II breast cancer patients is1.14, 95% CI: 0.68 – 2.00, P-value = 0.133. This indicates that patients in stage II are 1.14 times more likely to be at the

risk of death compared to patients in stage I (the reference stage), but the effect is not statistically significant at 1% level of significance.

For stage III patients, the hazard ratio is 2.87, 95% CI: 2.27 – 3.44, P-value 0.0001. The patients in stage III are 2.87 times more likely to be at the risk of death, compared to patients in stage I. This indicates that stage of the breast cancer have significant effect on the mortality of the patients at 1% level of significance.

The risk of death as a result of breast cancer increase as the patient gets older. The hazard ratio for age group ≤ 39 years is 0.61, 95% CI: 0.29 – 0.55, P-value = 0.044 Patients in age group ≤ 39 years are 0.61 times less likely to be at the risk of death, compared to patients in age group 60 years and above (reference group). 60 years and above is considered as reference group because age 60 and above are periods where the probability of death is high. The effect is not significant at 1% level of significance.

The hazard ratio of age group 40-49 years is 0.49, 95% CI: 0.59 – 0.89, P-value 0.027. The patients in this age group are 0.49 times less likely to be at the risk of death compared to patients in age group 60 and above years. The effect is not significant at 1% level of significance.

Similarly, patients in age group 50-59 years are 0.35 times less likely to be at the risk of death compared to patients in age group 60 and above years estimated HR = 0.35, 95% CI: 0.12 – 1.68, P-value = 0.174. The effect is not statistically significant at 1% level of significance.

4.6 Results from the Analysis of Weibull Model

Log likelihood = -209.30311

Table 4.7 Results for Weibull Model

Covariates	**Hazard Ratio (HR)**	**Std. Error**	**P-values**	**[95% Conf. Interval]**	
Occupation (H/W)	0.8577	0 .1648	0.0420	1 .5885	2.2502
Treat. (improved)	0.9144	0.2283	0.0020	1 .5606	1.4915
Stage II	1.1368	1.7425	0.0930	0. 8485	2.1653
Stage III	2.7763	1.2165	0.0035	2. 8674	3.2100
Age $\leq$ 39	0.5983	0.1791	0.0330	0.2463	0.4810
Age 40-49	0.4221	0.3775	0.0040	0.7324	1.5804
Age 50-59	0.2965	0.4349	0.2430	0.1625	1.7727

The hazard ratio computed for occupation is 0.86, 95% CI: 1.59 – 2.25, P-value = 0.0420. This indicates that housewives are 0.86 times less likely to be at the risk of death compared to patients in other occupations. Occupation of the patients does not have significant effect on the mortality of the patients at 1% level of significance.

The hazard ratio for results of the treatment is 0.91, 95% CI: 1.56 – 1.49, P-value = 0.0020. This indicates that breast cancer patients with improved condition are 0.76 less likely to be at the risk of death compared to patients with unimproved condition. This means condition of the patient is found to have significant effect on the mortality of the patients at 1% level of significance.

The hazard ratio for stage II breast cancer patients is1.14, 95% CI: 0.84 – 2.17,

P-value = 0.0930. This indicates that patients in stage II are 1.14 times more likely to be at the risk of death compared to patients in stage I (the reference stage). The effect is not statistically significant at 1% level of significance.

For stage III patients, the hazard ratio is 2.78, 95% CI: 2.87 – 3.21, P-value 0.0035. The patients in stage III are 2.87 times more likely to be at the risk of death, compared to patients in stage I. This revealed that stage of the breast cancer have significant effect on the mortality of the patients at 1% level of significance.

The risk of death as a result of breast cancer increase as the patient gets older. The hazard ratio for age group $\leq$ 39 years is 0.59, 95% CI: 0.25 – 0.48, P-value = 0.033 Patients in age group $\leq$ 39 years are 0.59 times less likely to be at the risk of death, compared to patients in age group 60 and above years (reference group). The effect is not statistically significance at 1% level of significance.

The hazard ratio for age group 40-49 years is 0.42, 95% CI: 0.73 – 1.58, P-value 0.004. The patients in this age group are 0.42 times less likely to be at the risk of death compared to patients in age group 60 and above years. The effect is statistically significant at 1% level of significance.

Similarly, patients in age group 50-59 years are 0.29 times less likely to be at the risk of death compared to patients in age group 60 and above years estimated HR = 0.29, 95% CI: 0.16 – 1.77, P-value = 0.243. But the effect is not statistically significant at 1% level of significance.

4.7 Results from the Analysis of Gompertz Model

Log likelihood = -213.1381

Table 4.8 Results for Gompertz Model

Covariates	**Hazard Ratio (HR)**	**Std. Error**	**P-value**	**[95% Conf. Interval]**	
Occupation (H/W)	0 .8542	0 .1621	0.0406	1 .5888	2.2391
Treat. (improved)	0.8685	0.2178	0.0254	1 .5313	2.4197
Stage II	1.1465	1.6452	0.1230	0. 8562	2.3523
Stage III	2.6941	1.4122	0.0035	2. 9641	3.3146
Age ≤ 39	0 .5695	0.2068	0.012 0	0 .2795	0.5031
Age 40-49	0.4338	0.3419	0.0080	0.9353	1.9724
Age 50-59	0.2867	0.4658	0.3210	0.2023	1.8944

The hazard ratio computed for occupation is 0.85, 95% CI: 1.59 – 2.24, P-value = 0.0406. This indicates that housewives are 0.85 times less likely to be at the risk of death compared to patients in other occupations. Occupation of the patients does not have significant effect on the mortality of the patients at 1% level of significance.

The hazard ratio for results of the treatment is 0.86, 95% CI: 1.53 – 2.41, P-value = 0.0254. This indicates that breast cancer patients with improved condition are 0.86 less likely to be at the risk of death compared to patients with unimproved condition. Condition of the patients does not have significant effect on the mortality of the patients at 1% level of significance.

The hazard ratio for stage II breast cancer patients is1.15, 95% CI: 0.86 – 2.35,

P-value = 0.123. This indicates that patients in stage II are 1.15 times more likely to be at the risk of death compared to patients in stage I (the reference stage). The effect is not statistically significant at 1% level of significance.

For stage III patients, the hazard ratio is 2.69, 95% CI: 2.96 – 3.31, P-value = 0.0035. The patients in stage III are 2.69 times more likely to be at the risk of death, compared to patients in stage I. The effect is statistically significant at 1% level of significance.

The risk of death as a result of breast cancer increased as the patient gets older. The hazard ratio for age group ≤ 39 years is 0.57, 95% CI: 0.28 – 0.50, P-value = 0.012 Patients in age group ≤ 39 years are 0.57 times less likely to be at the risk of death, compared to patients in age group 60 and above years (reference group). The effect is not statistically significant at 1% level of significance.

The hazard ratio in age group 40-49 years is 0.43, 95% CI: 0.93 – 1.97, P-value = 0.008. The patients in this age group are 0.43 times less likely to be at the risk of death compared to patients in age group 60 and above years. The effect is statistically significant at 1% level of significance.

The patients in age group 50-59 years are 0.29 times less likely to be at the risk of death compared to patients in age group 60 and above years estimated HR = 0.29, 95% CI: 0.20 – 1.89, P-value = 0.321. The effect is not statistically significant at 1% level of significance.

Table 4.9: Summary of the Results of the Analysis and Akaike Information Criterion

(AIC) for Cox, Weibull and Gompertz Regression Models

Covariates	Cox Model HR (95% Conf. Interval)	Weibull Model HR(95% Conf. Interval)	Gompertz Model HR(95% Conf. Interval)
Occupation	0.85 [1.56 - 2.24]	0.86 [1.59 - 2.25]	0.85 [1.59 - 2.24]
Results of the Treatment	0.76 [1.18 - 1.89]	0.91 [1.56 - 1.99]	0.87 [1.53 - 2.24]
Stage II of breast cancer	1.14 [0.68 - 2.00]	1.14 [0.85 - 2.17]	1.15 [0.86 - 2.35]
Stage III of breast cancer	2.87 [2.27 - 3.44]	2.78 [2.87 - 3.21]	2.69 [2.96 - 3.31]
Age at diagnose			
≤ 39years	0.61 [0.29 - 0.55]	0.59 [0.25 - 0.48]	0.57 [0.28 - 0.50]
40 -49 years	0.49 [0.59 - 1.89]	0.42 [0.73 - 1.58]	0.43 [0.94 - 1.97]
50 - 59 years	0.35 [0.35 - 1.68]	0.29 [0.16 - 1.77]	0.27 [0.20 - 1.89]
Akaike Information Criterion (AIC)	530.65	422.60	430.28

This research work compared estimation methods using parametric (Weibull and Gompertz) and semi-parametric Cox models under proportional hazards framework of survival data. Overall model comparison was done using Akaike Information Criterion (AIC). As observed, from Table 4.9 Weibull regression model had the least AIC value which shows best performance in handling breast cancer data, Cox regression model with highest AIC value performed less. Our results agreed with the results obtained by Hui., *et al* (2011) where they compared Cox and Weibull models using gastric cancer data.

They concluded that their data supported Weibull model as an alternative to Cox model based on AIC.

On checking the proportionality assumption for the four categorical covariates used in the analysis, it was observed that the assumption was valid for stage of the disease and the results of the treatment, where as, it was not valid for age and occupation. Cox proportional hazard is widely used in analysing survival data in clinical research. Altman, *et al* (1985) found that only five percent of all the studies using Cox models check the assumption. However, Cox proportional hazard assumption is not always satisfied in the data. If this assumption does not hold, there are various solutions to consider. One solution is to include time – dependent variable for the predictors with non-proportional hazards.

This approach is used to account for a variable with non – proportionality, different results may be obtained from different choices of time dependent variables. Alternatively, we can use a model where we can stratify on proportional predictors. The stratified Cox model is not appropriate when the covariate with non-proportionality is continuous or of direct interests. Both ways are still based on the comparison of hazards.

According to Ravangard, *et al* (2011), if the assumption of proportional hazards is not fulfilled, the reliability of the results of the Cox proportional hazards is low and it would be preferred to use parametric models.

The Accelerated Time Failure (AFT) model is an alternative method for the analysis of survival data even when hazards are not proportional. Based on asymptotic results, the AFT models should lead to more efficient parameter estimates than Cox model under certain circumstances (Cox and Oakes, 1984, Oakes, 1977).

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary

With Kaplan-Meier survival analysis procedure, the study examined the distribution of time of the breast cancer patients. The mean age of the patients was found to be 43.39 years with the standard deviation of 11.74 years. The overall median survival time was 10 months. The 5-year overall survival rate was about 35.0%. The median survival time for housewives was 16 months and for other occupations was 10 months. Median survival time for age group $\leq$ 39 years was 19 months and for age group 40-49, 50 -59 and 60 years and above were 7, 5 and 10 months respectively.

The Log-rank tests were used to compare the survival curves of the breast cancer patients. For stages of the breast cancer patients, the Log-rank statistic is 300.23 with P-value of 0.001. For occupation of the breast cancer patients the Log-rank statistic is 0.01 with P-value of 0.9899. In the case of age groups, the Log-rank statistic is 8.81 with P-value of 0.0378. Lastly, for the results of the treatment, the Log-rank statistic is 139.11 with P-value of 0.0001.

The Cox model estimates revealed that the hazard ratio computed for occupation is 0.85. Hazard ratio for results of the treatment is 0.76. The hazard ratio for stage II breast cancer patients' is1.14. The hazard ratio for stage III breast cancer patients is 2.87. The estimated hazard ratio for age group $\leq$ 39, 40 – 49, 50 – 59, 60 and above years are 0.61, 0.49 and 0.35 respectively.

For Weibull model estimates, the hazard ratio for occupation is 0.85. The hazard ratio for stage II patients is 1.14. Similarly, hazard ratio for stage III patients is 2.78.

But hazard ratio for the results of the treatment is 0.91. The hazard ratio for age group $\leq$ 39, 40 – 49, 50 – 59 and 60 and above years are 0.50, 0.42 and 0.29 respectively.

The Gompertz regression model estimates of hazard ratio for occupation is found to be 0.854. The hazard ratio for results of the treatment is 0.869. The hazard ratio for stage II patients is 1.15. The hazard ratio for stage III breast cancer is 2.69.

The hazard ratio for age groups $\leq$ 39, 40 – 49, 50 – 59 and 60 and above years are 0.57, 0.423 and 0.29 respectively.

5.2 Conclusion

In this study, Kaplan-Meier procedure was used to estimate the survival curves of breast cancer patients at Ahmadu Bello University Teaching Hospital Zaria. The mean age of the patients was found to be 43.39 years with the standard deviation of 11.74 years, and overall median survival time of 10 months. This indicates that 50% of the breast cancer patients survived less or equal to 10 months and the other 50% survived beyond 10 months after they were diagnosed with the disease. This is the survival time at which the cumulative survival function is equal to 0.5. The 5-year overall survival rate of about 35.0% of breast cancer patients admitted at Ahmadu Bello University Zaria, between 1997 and 2012 is low compared to survival rates of some developing nations.

The Log-rank statistic for stages of the breast cancer patients is 300.23 with P-value of 0.0001; this indicates that the survival experience of the breast cancer patients are not the same with respect to the stages of breast cancer at 1% level of significance.

Log-rank statistic for age groups is 8.81 with P-value of 0.0378; this indicates that the four age groups have the same survival experience at 1% level of significance. The log-rank statistic for the results of the treatment is 139.11 with P-value of 0.0001; this indicates

that breast cancer patients have significant different survival experience with regard to the results of their treatment at 1%. On the other hand, Log-rank statistic for occupation of the breast cancer patients is 0.01with P-value of 0.9899, this result indicates that the survival curves for housewives and other occupations are the same at 1% level of significance.

From the results of the analysis obtained, for Cox, Weibull and Gompertz regression models, occupation of the breast cancer patients, stages of the cancer disease and results of the treatment have significant effect on the mortality of the patients, and increase in the risk of death as a result of breast cancer is associated with increase in age.

This research work analysed breast cancer dataset using three proportional hazards models, Cox model under semi-parametric model framework and also Weibull and Gompertz under parametric model frame work. The results of the analysis revealed that, according to our breast cancer data, the parametric Weibull regression model performed best and Cox regression model performed less.

The results of this study shows that, according to our breast cancer data, the parametric Weibull regression model could better determine the factors associated with the beast cancer disease than the semi-parametric Cox proportional hazards and parametric Gompertz models. In other words, in the present study, the Weibull model provided a better fit to the study data than Cox proportional hazards and Gompertz models. Therefore, it would be better for researchers of the health care field to consider this model in their researches concerning the breast cancer disease, if the assumption of proportional hazards is not fulfilled.

5. 3 Recommendations

The following are the recommendations made:

- Public awareness regarding early detection of breast cancer is recommended.
- Necessary strategies to enhance cancer registry for breast cancer in an effort to increase the level of case ascertainment and completeness of the data is recommended.
- Further research on survival rate is recommended to observe the temporal changes in survival rate among Nigerian breast cancer patients which reflect the impact of the effectiveness of prevention programmes.
- The federal government should provide more funds so that researchers can make discoveries which will significantly enhance progress in the understanding of the epidemiology of the disease.
- Further research should also be made by the government and health practitioners on new drugs that will effectively combat breast cancer disease.

5.4 Suggestions for Further Research

This study can be extended using the methods of Cox, Weibull and Gompertz regressions under different samples size and percentage of censoring. This study can also be extended by comparing these models with Logistic regression model to find out which of the models can better fit the data and give the best result.

REFERENCES

Aggrey, S. E. (2002). Comparison of three Nonlinear and Spline Regression Models for describing Chicken Growth Curves. *Poultry Sci.* **81**, 1782–1788.

Agresti A. (2002). *Categorical Data Analysis*.(2nd ed) New York: John Wiley and Sons.

Ahuja J.C. and Nash S.W. (1967). The Generalised Gompertz-Verhulst family of Distributions. *Sankhya Ser A*; 29(2): 141 - 156

Akaike H. (1973). "Information Theory and Extension of the Maximum Likelihood Principle ".In B.N. Petrov and F. csaki (eds) 2nd *International Symposium on Information Theory, Academia Kiodo, Budapest* 267 – 281

Altman D.G. and de Stavola B. L. (1994). *Practical Problem in Fitting Proportional Hazard Model to Data with Updated Measurement of Observational Studie*s. John Wiley and Sons: Chicester.

Altman D.G., (1977). *Practical Statistics for Medical Research*. Chapman and Hall/CRC: Boca Raton.

Bang J. W. (1949). Maximum Likelihood Estimates of Proportion of Patients. Curd by Cancer Therapy. *Journal of Royal Statistical Society Series B.*

Berry G. (1983). The Analysis of Mortality by the Subject Year Method. *Biometrics* 30: 173 – 184

Berwick V., Cheek L., Bell T. (2004). Statistical Review 12: *Survival Analysis Crit. Core*. 8: 389 – 394

Bijan Moghimi-Dehkordi, Azdeh Safee, Mohamad Amin Pourhoseigholi, Reza Fatemi, Ziaoddi Tabeie, Mohammed Reza Zali (2008) Statistical Comparison of Survival Models for Analysis of Cancer Data. *Asian Pacific Journal of Cancer Prevention* Vol. 9

Blossfield H. P., Hamerle A., Mayer K.U. (1989). *Event History Analysis*. Hall Dale, N. J. Lawrence Erlbaum

Borkowf C. R. (2005). A Sample Hybrid Variance Estimator for the Kaplan – Meier Functions. *Statistics in Medicine* 24: 827 - 851

Breslow N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics,* 30: 89 - 99

Breslow N. (1990).Testing of Hypothesis in Over – Dispersed Poisson Regression and Other Quasi *Stat. Associ*. 85: 565 – 571

Bryan and David P. (1974). Selecting Optimal Treatment in Clinical Trials Using Covariates Information. *Annual Meeting of American Statistical Association*. Aug. 28

Cameron A. C. and Famoye F. (1992). *Regression Analysis of Count Data*. Cambridge University Press, New York.

Cameron A. C. and Trivedi P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, New York

Cameron A.C., Trivedi P. K., Milue F., Piggott J. (1988). *A Micro-Econometrics Model of Demand for Health Cars and Health Insurance Review of Economic Studies* 55: 85 – 106

Cohen A. C. and Whitten B. (1982). Modified Maximum Likelihood and Modified Moment Estimators for Three Parameters Weibull distribution. *Comm. Stat. Theory Methods II:* 2631 – 2656

Cohen, Jacob, Patricia, West, Steven, Aikens, Leonas (2002). *Applied Multiple Regression/ Correlation Analysis for the Behavioural Science* (3rd ed.) Rout ledge

Collett D. (1994). *Modeling Survival Data in Medical Research* Chapman and Hall: London.

Consul C. and Famoye F. (1992). Generalised Poisson Regression Model. *Communication in Statistics.*

Cornfield J., Gordon T., Smith W.N. (1961). Quantal Response Curves for Exponential Uncontrolled Variables *Bull Int. Inst*. 38: 97 – 115

Cox D. R. (1975). Partial Likelihood. *Biometrika* 62: 269 – 276

Cox D. R. and Oakes D. (1984). *Analysis of Survival Data New York, NY* Chapman and Hall Ltd

Cox D.R. (1972). Regression Models and Life Tables. *Journal of Royal Statistical Society Series B* 34: 187 - 220

Crawley M. J. (2007). *The R Book,* Wiley J. and Sons Chichester UK: pp 527.

Davis D. J. (1952). An Analysis of Failure Data. *Journal of American Statistical Association 47*: 113 – 150

Economos A. (1982) Rates of aging, Rate of Dying and Mechanism of Mortality. *Archives of Gerontology and Gariatrics*, 1 (1)

Efron B. (1977). The Efficiency of Cox Likelihood Function for Censored Data. *Journal of American Statistical Association 72*: 557 – 565

Epstein B. (1958) *Exponential Distribution and its Role in Life Testing*. Industrial quality Control 15: 2 -7

Elvan A. H., Asli S., Burak U., Omar D., Mehmet N. O., Gul K. (2010). Comparison of Five Survival Models: Breast Cancer Registry Data from Ege University Cancer Registry Center. *Turkiye Klinikleri J. Med. Sci. 30 (5).*

Feinleib M. and Macmahan B. (1960). *Variation in Duration of Survival of Patients with Chronic Leukemia:* 15(3): 332 – 349

Gaddum J. H (1945). *Log-normal Distribution, Nature Lon*don: 156 (3964): 463 – 466

Gardener W., Mulvey E.P., Show E.C. (1995). Regression Analysis of Count and Rates: Poisson and Negative Binomial Models : *Psychological Bulletin,* 118: 394 – 404

Gehan E. A. (1965). A Generalised Wilcoxon Test for Comparing Arbitrary Single – Censored Samples. *Biometrica 52*: 203 - 223

Gill R. D. (1980). Censoring and Stochastic Integrals, Volume 24, Mathematical Centre Tracts. Amsterdam: *Mathematisch Centrum.*

Gokovali U., Bahar O., Kozak M. (2007). Determinant of Length of Stay: A Practical Use of Survival Analysis. *Tourism Management 28(3),* 736 - 746

Golub J. (2007). Survival Analysis and European Union Decision Making. *European Union Politics 8(2),* 155 – 17

Gompertz, B. (1825). "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies". *Philosophical Transactions of the Royal Society of London* **115**: 513–583.

Greenwood M. (1926). A Report on Natural Duration of Cancer. *In Report on Public Health and Medical Subjects Vol. 23 His Majesty's Stationery Office London*. 1- 26

Harirchi I., Kalahdoozan S. and Karbakhsh M. (2001). Twenty Years of Cancer in Iran: Downstaging without a formal screening program, *Ann Oncol*., 22, 93 – 97.

Harrel F., Lee K, Mark D.(1996). Multivariable prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequecy Measuring and Reducing Errors: *Statistics in Medicine 15*: 361 – 387

Homer R. D. (1987). Age at Onset of Alzheimer 's disease, Due to Relative Importance of *Etiologic Factor 126(3):* 409 – 414

Hosmer D. W. and Lemeshow S. (1998). *Applied Survival Analysis*: Regression Modeling of Time to Event Data. Wiley New York Inter Science.

Hosmer, David W., Lemeshow, Stanley (2000). *Applied Logistic Regression* (2nd ed.) Wiley ISBN0 - 471 – 35632 - 8

Hossaini A. M. and Zimmer W.J. (2003). Comparison of Estimation Methods for Weibull Parameters: Complete and Censored Samples: *Journal of Statistics Compt. Simulation 73:* 145 – 153

Hui P. Z., Xin X., Chuan H. Y., Ahmed A., Shun F. L., Yu K. D., (2011). Application of Weibull Regression Model for Survival of Patients with Gastric Cancer (3rd ed.) *BioMed Central* John Wiley.

Inhelmann R., Grande E., francisci E., De Angelis R., Micheli A. , Verdecchic A. (2005). *National Estimate of Cancer Patient's Survival in Italy a Model based Method*.

Jaheen Z. F. (2003). Prediction of Progressive Censored Data from the Gompertz Model. *Communication in Statistics-Simulation and Computation; 32(3):* 663 -676

Jiezhi Q. (2009). Comparison of Proportional Hazards and Accelerated Failure Time Models. An M.Sc. Thesis.

Johansen S. (1978). The Product Limit Estimate as a Maximum Likelihood Estimate. *Scandinavian Journal of Statistics 5:* 195 – 199

Kalbleisch J.D. and Patience R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley New York.

Kaplan E. L. and Meier P. L. (1958). Non Parametric Estimation from Incomplete Observations. *Journal of American Statistical Association 53*: 457 – 481

Klein J. P. and Moeschberger M. L. (1997). Survival Analysis Techniques for Censored and Truncated Data, *Springer – Verlog*, New York.

Kleinbaum D. G. and Klien M. (2005). *Survival Analysis: A Self – Learning Text* 2nd edition *USA*. Springer.

Lee E. T. (1980). *Statistical Methods for Survival Data Analysis Life Time Learning*: Publication Belmont California.

Lee E. T. and Wang J. W. (2003). *Some Well – Known Parametric Survival Distributions and their Application.* New Jersey: John Wiley and Sons. Inc. 134 – 162

Lee H. P. (1982). On Clinical Trials and Survival Analysis. Singapore *Medical Journal 23:* 164 - 167

Mantel N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration. *Cancer Chemotherapy Report 50*: 163 – 170

Marubini E. and Valsecchi M. D. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies.* John Wiley and Sons: Chichester.

Medawar, P.B., (1940). Growth, growth energy and ageing of the Chicken's Heart. *Proc. Royal Soc. London,129*: 332 - 355.

Mohamed A. P., Ebrahim H., Bijan M. D., Azadi S., Alireza A., Mohammed R.Z. (2007). Comparing Cox Regression and Parametric Models for Survival of Patients with Gatric Carcinoma. *Asian Pacific Journal of Cancer Prevention. Vol. 8*

Nor A. A., Wan R.W., Asiah M. N., Ali Z., Ibrahim S. (2013) Survival Rate of Breast Cancer Patients in Malaysia. *Asian Pacific Journal of Cancer Prevention Vol.14.*

Oakes D. (1977). The Asymptotic Information in Censored Survival Data. *Biometrika 64:* 441-448

Olshansky S. J. and Carnes B. A. (1997). Ever since Gompetrz. *Demography 34:* 1 -15

Palei S. K. and Das S. K. (2009). Logistic Regression Model for Prediction of Roof full Risks in Board and Pillar Working in Coal Mines: *An Approach Safety Science 47*: pp 88

Parodi S, and Battarelli (2006). Poisson Regression Model in Epidemiology An Introduction. *Ann Fac. Medic. Deparma* 25 – 44

Peto R., Pike M. C., Armitage P., Breslow N. E. Cox D. R., Howard S. V., Mantel N. McPhersonK., Peto J., Smith P. G. (1977). Design and Randomized Clinical Trials Requiring Prolong Observation of each Patient. *British Journal of Cancer, 35*: 1 – 39

Ravangard R., Mohamed A., Rashidian A., Ali A., Zare A. and Hjjat Z. (2011). Comparison of the Results of Cox PH Models and Parametric Models in the Study of Length of Stay in a Tertiary Teaching Hospital in Iran. *Acta Medica Iranica; 49(10)* : 650 - 658

Roche J. R. and Berry D. P. (2006). Per – Parturient Climatic Animal and Management Factors Influencing the Incidence of Milk Fever in Grazing Systems: *Journal of Dairy Science, 89:* 2775 – 2783

Rose N. (1990). Profitability and Product Quality: Economic Determinant of Airline Safety Performance, *Journal of Political Economy 98*: 944 - 964

Sariego J. (2010).Breast Cancer in the Young Patients: *The American Surgeon 76(12):*1397 – 1401

Saunders, C. S. (2007) Reliability, Life Testing and the Prediction of service Lives: For Engineers and Scientists. Springer, New York.

Singh K. P., Wulu J. T. Bartolucci A. A. (2001). *A Note on Generalised Poisson Regression Models* at KSingh@hsc.unit.edu.jwutu@hrsa.gov

Sloot T. and Verschuren (1990). Decision- Making Speed in European Community. *Journal of Common Market Studies. 29(1)* 75 – 85

Spiegelhalter D. J. (1986). Probabilistic Prediction in Patient Management and Clinical Trials. *Statistics in Medicine 5*: 421 – 433

Stephen J. W. (2009). *What is a Cox Model?* 2nd ed. What is …? Series. Heyward Group Ltd. UK.

Stewart B. W. and Kleihues P. (2003). *The Global Burden of Cancer World Cancer Report,* (1st ed.) Lyon IARC Press.

Talback M., Stenbeck M., and Rossen M. (2004). Up-to-date Long Term Survival of Cancer Patients: An Evaluation of Period Analysis on Swedish Cancer Registry Data *Eur. J. Cancer, 40:* 1361 – 1371

Trvier N., Gridley G., Blair A., Dosmeci M., Boffetta P. (2003). Cancer Incidence Among Males Swedish Veterinarians and other Workers of Veterinary Industries: *A Record – Linkage Study Cancer Causes and Control 14:* 587 – 598

Van Groenedael (2010).Application of Kaplan - Meier Estimation on Cancer Data. *Biometrica 27:* 20 - 24

Visconi S., Pastora G., Dana E. Zuccola L. Peace N. Marleti F. (2006). Life Expectancy as an Indicator of Outcome in Follow up Population Based Cancer Registries: *Ann Oncol 17(1)* 167 – 171

Walter S. and Duncan D. (1967). Estimation of Probability of Event as a Function of Several Variables, *Biometrika 54*: 167 – 179

Weibull W. A. (1951). Statistical Distribution of Wide Applicability*: Journal of Applied Mechanics 18*: 293 – 297

Wingo P.A., Tong T. and olden S. (1995). Cancer Statistics C. A. : *Cancer J. Clin. 46(1),* 5- 27 doi.

Winkelman R. (1995). Duration Dependency and Dispersion in Count Data Models: *Journal of Business and economics Statistics, 13:* 467 – 474

Winsor, C. P. (1932). The Gompertz curve as a growth curve. *P. Natl. Acad. Sci. USA* **18**, 1–8.

World Health Organisation (WHO) *Reports (1990*)

World Health Organisation (WHO) *Reports (2008)*

Wright, S. (1986). *Book Review of* the Biology of Population Growth by Raymond Pearl and the Natural Increase of Mankind by J. Shirley Sweeney. J. *Am. Stat. Assoc.* **21**, 493–497.

Wulu J.J., Singh K.P., Famoye F. McGwin (2002). Regression Analysis of Count Data. *Journal of Indian Society of Agriculture Statistics, 55*: 220 – 231.

Zhao G. (1996). The Homogenic stimate for the Variance of Survival Rates. *Statistics in Medicine, 24*: 827 – 851

Zolen M. (1966). The Application of Exponential Model to Problem in Cancer Research, *Journal of Royal Statistical Society Series, 3*: 368 - 398

APPENDIX I

Raw data of the patients collected from Ahmadu Bello University Teaching Hospital Zaria.

S/n	Age	Sex	Occupation	Stage of disease	Length of Stay (in days)	Status	Results of Treatment
1	2	0	1	I	28	1	0
2	3	0	1	I	98	1	0
3	2	0	1	I	14	1	0
4	1	0	1	III	22	0	0
5	1	0	1	III	10	0	0
6	1	0	1	I	78	1	0
7	1	0	1	I	28	1	0
8	4	0	1	I	22	1	0
9	1	0	1	II	2	1	1
10	1	0	1	II	25	1	1
11	3	0	1	II	15	1	1
12	2	0	1	II	50	1	1
13	3	0	1	II	2	1	1
14	2	0	1	II	21	1	1
15	2	0	2	III	15	0	0
16	5	1	3	II	34	1	1
17	2	0	1	II	34	1	1
18	2	0	1	II	57	1	1
19	2	0	1	III	25	0	0
20	3	0	1	II	20	1	0
21	1	0	1	II	5	1	1
22	1	0	2	II	42	1	1
23	2	0	3	II	31	1	0
24	2	0	1	I	18	1	0
25	5	0	1	I	6	1	0
26	3	0	1	I	50	1	0
27	4	0	1	I	98	1	0
28	2	0	1	I	14	1	0
29	2	0	2	II	38	1	1
30	2	0	1	II	13	1	1
31	3	0	1	III	13	0	0
32	2	0	1	III	14	0	0
33	2	0	1	II	40	1	1
34	4	0	1	III	3	0	0
35	1	0	2	II	5	1	1
36	2	0	1	II	2	1	1
37	1	0	1	II	9	0	0
38	4	0	1	II	15	1	1
39	1	0	1	II	4	1	1
40	1	0	1	II	23	1	1
41	1	0	1	II	1	1	0

42	1	0	1	II	35	1	1
43	2	0	1	II	32	1	1
44	2	0	1	III	1	0	0
45	2	0	1	II	11	1	1
46	2	0	1	II	32	1	1
47	3	0	1	II	44	1	1
48	3	0	1	II	8	1	0
49	2	0	1	II	30	1	0
50	3	0	1	II	19	1	0
51	2	0	1	II	20	1	1
52	3	0	1	II	36	1	1
53	3	0	1	II	8	1	1
54	3	0	1	II	17	1	1
55	3	0	2	II	12	1	1
56	3	0	6	II	19	1	1
57	4	0	1	II	22	1	1
58	2	0	4	II	20	1	1
59	5	0	1	II	14	1	1
60	2	0	1	II	13	1	1
61	1	0	1	II	14	1	1
62	2	0	1	II	21	1	1
63	3	0	1	II	22	1	1
64	3	0	1	II	30	1	1
65	3	0	2	II	16	1	1
66	5	0	1	II	17	1	1
67	3	0	1	II	18	1	1
68	3	0	1	II	19	1	1
69	3	0	1	II	22	1	1
70	5	0	1	II	20	1	1
71	2	0	1	II	13	1	1
72	3	0	6	II	14	1	1
73	1	0	1	II	19	1	1
74	5	0	1	II	29	1	1
75	3	0	6	II	21	1	1
76	1	0	6	II	22	1	1
77	6	0	6	II	21	1	1
78	2	0	1	II	13	1	1
79	6	0	1	II	19	1	1
80	6	0	3	II	20	1	1
81	3	0	6	II	11	1	1
82	2	1	6	II	16	1	1
83	6	0	1	II	19	1	1
84	3	0	4	II	20	1	1
85	3	0	1	II	19	1	1

86	5	0	1	II	19	1	1
87	3	0	4	II	11	1	1
88	3	0	6	II	17	1	1
89	1	0	4	II	19	1	1
90	3	0	1	II	2	1	1
91	2	0	1	II	2	1	1
92	2	0	1	II	2	1	1
93	2	0	6	II	10	1	1
94	2	0	1	III	1	0	0
95	4	0	1	III	10	0	0
96	3	0	1	II	2	1	1
97	3	0	1	III	2	0	0
98	1	0	1	III	3	0	0
99	2	0	1	III	10	0	0
100	2	0	4	III	4	0	0
101	4	0	4	III	3	0	0
102	3	0	1	III	6	0	0
103	4	0	1	III	5	0	0
104	2	0	1	III	7	0	0
105	1	0	3	II	1	1	1
106	1	0	1	II	2	1	1
107	3	0	1	II	5	1	1
108	2	0	1	II	3	1	1
109	2	0	1	II	2	1	1
110	3	0	1	III	2	0	0
111	2	0	1	II	2	1	1
112	3	1	6	II	2	1	1
113	2	0	1	II	3	1	1
114	3	0	1	III	4	0	0
115	3	0	1	III	5	0	0
116	1	0	1	III	3	0	0
117	3	0	1	III	5	0	0
118	1	0	1	III	5	0	0
119	3	1	2	III	2	0	0
120	3	0	1	III	2	0	0
121	3	0	1	III	2	0	0
122	1	0	6	III	6	0	0
123	1	1	2	III	3	0	0
124	4	0	1	III	3	0	0
125	2	0	2	III	2	0	0
126	4	0	6	III	2	0	0
127	1	0	1	III	10	0	0
128	2	0	1	III	3	0	0
129	2	0	1	III	3	0	0

130	3	0	1	III	2	0	0
131	2	0	1	III	3	0	0
132	2	0	5	III	1	0	0
133	4	0	1	II	2	1	1
134	2	0	1	III	2	0	0
135	5	0	6	III	2	0	0
136	3	0	1	III	10	0	0
137	1	0	1	II	3	1	1
138	4	0	1	III	2	0	0
139	2	0	1	III	2	0	0
140	3	0	1	III	1	0	0
141	3	0	2	II	3	1	1
142	2	0	1	III	4	1	1
143	2	0	1	III	2	0	0
144	2	0	1	III	3	0	0
145	3	0	1	III	5	0	0
146	3	0	1	III	5	0	0
147	2	0	1	III	6	0	0
148	2	0	1	II	6	1	1
149	2	0	1	III	24	0	0
150	4	0	1	III	1	0	0
151	5	0	1	III	7	0	0
152	2	0	4	II	8	1	1
153	4	1	2	III	2	0	0
154	4	0	1	III	2	0	0
155	5	0	1	III	2	0	0
156	3	0	1	III	2	0	0
157	2	0	1	III	2	0	0
158	4	0	1	II	3	1	1
159	2	0	4	II	2	1	1
160	3	0	1	II	3	1	1
161	3	0	1	II	8	1	1
162	4	0	1	III	2	0	0
163	4	0	6	III	3	0	0
164	1	1	2	III	2	0	0
165	4	0	1	III	2	0	0
166	2	0	1	III	3	0	0
167	1	0	1	III	2	0	0
168	2	0	1	III	2	0	0
169	3	0	1	III	2	0	0
170	2	0	1	III	2	0	0
171	2	0	1	II	4	1	1
172	4	0	1	III	2	0	0
173	2	0	1	III	9	0	0

174	4	0	1	III	3	0	0
175	2	0	1	II	2	1	0
176	3	0	1	II	2	1	1
177	4	1	3	II	2	1	1
178	3	0	2	III	2	0	0
179	2	0	1	III	2	0	0
180	5	0	1	III	2	0	0
181	3	0	1	II	2	1	1
182	2	0	2	III	2	0	0
183	5	0	1	II	2	1	1
184	2	0	1	III	2	0	0
185	1	0	1	II	2	1	1
186	4	0	1	II	2	1	1
187	3	0	1	III	2	0	0
188	3	0	1	III	2	0	0
189	3	0	2	II	2	1	1
190	1	0	5	III	4	0	0
191	3	0	1	III	2	0	0
192	2	0	1	III	2	0	0
193	2	0	2	III	2	0	0
194	4	1	3	III	2	0	0
195	4	0	1	III	2	0	0
196	1	0	6	II	2	1	0
197	3	0	1	III	6	0	0
198	2	0	1	III	5	0	0
199	2	0	1	III	4	0	0
200	3	0	4	III	6	0	0
201	3	0	4	II	1	1	0
202	3	0	1	II	3	1	0
203	2	0	1	II	4	1	0
204	4	0	1	II	4	1	0
205	1	0	1	II	2	1	0
206	3	0	1	II	2	1	0
297	4	0	2	III	2	0	0
208	4	0	1	III	2	0	0
209	2	0	2	III	2	0	0
210	2	0	1	III	3	0	0
211	3	0	2	III	5	0	0
212	3	0	2	II	3	1	0
213	4	0	1	III	2	0	0
214	1	0	1	II	6	1	1
215	1	0	1	II	5	1	1
216	3	0	1	II	2	1	1
217	3	1	3	II	2	1	0

218	2	0	1	II	2	1	1
219	3	0	2	II	3	1	1
220	5	0	6	II	2	1	0
221	4	0	1	II	2	1	0
222	3	0	1	III	2	0	0
223	2	0	1	III	2	0	0
224	2	0	1	III	2	0	0
225	3	0	6	III	2	0	0
226	2	0	2	III	2	0	0
227	4	1	6	III	2	0	0
228	1	0	1	III	2	0	0
229	3	0	1	III	2	0	0
230	1	0	1	III	3	0	0
231	5	0	1	III	2	0	0
232	4	0	2	III	2	0	0
233	4	0	1	III	3	0	0
234	3	1	3	III	3	0	0
235	3	0	1	III	4	0	0
236	5	0	1	III	2	0	0
237	3	0	1	III	2	0	0
238	2	0	2	III	2	0	0
239	3	0	1	II	3	1	1
240	6	0	1	II	2	1	1
241	2	0	4	II	17	1	1
242	1	0	6	II	64	1	1
243	1	0	1	III	22	1	1
244	1	0	1	II	19	0	0
245	3	0	1	II	17	1	1
246	2	0	1	II	18	1	1
247	3	0	6	II	13	1	1
248	1	0	1	II	11	1	1
249	2	0	1	II	12	1	1
250	1	0	6	III	23	0	0
251	1	0	1	II	1	1	1
252	3	0	1	II	3	1	1
253	3	0	1	II	4	1	1
254	1	0	5	III	2	0	0
255	1	0	2	III	2	0	0
256	3	0	1	III	2	0	0
257	3	0	1	III	9	0	0
258	3	0	6	III	2	0	0
259	1	0	5	III	36	0	0
260	2	0	1	III	3	0	0
261	3	0	1	II	4	1	1

262	4	0	6	III	2	0	0
263	3	0	1	III	2	0	0
264	4	0	6	III	1	0	0
265	4	0	6	III	2	0	0
266	2	0	1	III	21	0	0
267	3	0	1	III	4	0	0
268	1	0	1	III	5	0	0
269	2	0	1	III	2	0	0
270	1	0	1	III	2	0	0
271	1	0	1	III	2	0	0
272	3	0	1	III	2	0	0
273	3	0	4	II	2	1	1
274	4	0	1	II	4	1	1
275	1	0	1	III	10	0	0
276	5	1	4	III	2	0	0
277	3	0	4	III	2	0	0
278	2	0	1	III	2	0	0
279	1	0	1	III	3	0	0
280	2	0	1	III	3	0	0
281	5	0	1	III	3	0	0
282	2	0	1	III	2	0	0
283	3	0	1	III	2	0	0
284	2	0	2	III	6	0	0
285	3	0	1	II	2	1	1
286	3	0	2	II	2	0	0
287	3	0	1	III	2	0	0
288	2	0	1	III	2	0	0
289	1	0	1	III	2	0	0
290	4	0	1	III	2	0	0
291	3	0	1	II	5	1	1
292	2	0	5	II	8	1	1
293	2	0	5	II	10	1	1
294	5	0	1	III	10	0	0
295	2	0	4	III	6	0	0
296	4	0	1	III	7	0	0
297	1	0	5	II	11	1	1
298	5	0	4	III	16	0	0
299	2	0	6	II	18	1	1
300	4	0	2	II	17	1	1
301	2	0	2	II	18	1	1
302	3	0	1	III	5	0	0
303	2	0	1	II	4	1	1
304	2	0	1	II	7	1	1
305	2	0	5	II	8	1	0

306	4	0	1	II	4	1	1
307	3	0	4	II	5	1	0
308	3	0	1	II	3	1	1
309	3	0	1	II	6	1	1
310	3	0	1	II	3	1	1
311	3	0	1	II	5	1	1
312	3	0	1	II	7	1	1