3-26-2016

Modeling and Survival Analysis of Breast Cancer: A Statistical, Artificial Neural Network, and Decision Tree Approach

Venkateswara Rao Mudunuru

Modeling and Survival Analysis of Breast Cancer:

A Statistical, Artificial Neural Network, and Decision Tree Approach

by

Venkateswara Rao Mudunuru

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics and Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Lesław A. Skrzypek, Ph.D.
Gangaram S. Ladde, Ph.D.
Yuncheng You, Ph.D.
Marcus McWaters, Ph.D.

Date of Approval:
March 23, 2016

Keywords: Statistical Modeling, Survival Analysis, Parametric Analysis, Probability Distribution, Decision Trees, Artificial Neural Networks, Classification.

DEDICATION

I dedicate this dissertation to my father Raghunadha Rao Mudunuru, my mother Parvathi Devi, my wife Vidya Bhargavi Mudunuru and our cute daughter Vishnu Kruti. Their loving words of encouragement from the depths of his magnanimous heart and care have solely contributed to my commitment to complete the doctoral degree.

I would also like to dedicate this dissertation to my mentor and my role model the late honorable Dr. V. Lakshmikantham of Florida Institute of Technology, Melbourne, FL, who is the reason for what I am today.

ACKNOWLEDGEMENTS

The real spirit of achieving a goal successfully is through the way of excellence and dedicated discipline. Without the cooperation, encouragement and the help provided to me by various personalities, I would have never accomplished this task. I thank one and all who contributed to the realization of this dream.

First of all, I render my gratitude to the Almighty who bestowed self-confidence, ability and strength in me to complete this work. Without his grace this would never come to be today's reality.

With deep sense of gratitude I express my sincere thanks to my esteemed supervisor, University Professor Dr. Lesław A. Skrzypek, in carrying out this work under his effective supervision, great help and support during the last three years.

Thanks are also due to my committee members, Professors Dr. Gangaram Ladde, Dr. Yuncheng You and Dr. Marcus McWaters for providing his unconditional support, cooperation in completing this dissertation and for sharing his ideas about how to grow in the research environment. I shall be failing in my duties if I do not express my deep sense of gratitude towards Dr. Chris P Tsokos, who helped me to quickly adjust to the new study environment in this completely new country.

My personal thanks to Dr. Rajaram Lakshminarayan, Affiliate Professor, College of Public Health, and Dr. Rebecca Wooten, Assistant Professor, Department of Mathematics and Statistics, for their constant source of inspiration that was a great importance in the completion of this dissertation.

In addition, my sincere thanks and love to all my friends Sampath Kalluri, Jayadeva Sai Sravan Patchava, Shashank Kanna, Patrick Assonken, Nana Bonsu, Neranga Fernando, Zahra Kottabi, and Keshav Pokhrel for their unconditional support and concern.

I am also thankful to all the staff members of the Department of Mathematics and Statistics for their full cooperation and help.

My greatest thanks are to all who wished me success especially my parents, my wife and my friends whose support and care made this dream come true.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Survival analysis today is widely implemented in the fields of medical and biological sciences, social sciences, econometrics, and engineering. The basic principle behind the survival analysis implies to a statistical approach designed to take into account the amount of time utilized for a study period, or the study of time between entry into observation and a subsequent event. The event of interest pertains to death and the analysis consists of following the subject until death. Events or outcomes are defined by a transition from one discrete state to another at an instantaneous moment in time. In the recent years, research in the area of survival analysis has increased greatly because of its large usage in areas related to biosciences and the pharmaceutical studies. After identifying the probability density function that best characterizes the tumors and survival times of breast cancer women, one purpose of this research is to compare the efficiency between competing estimators of the survival function. Our study includes evaluation of parametric, semi-parametric and nonparametric analysis of probability survival models.

Artificial Neural Networks (ANNs), recently applied to a number of clinical, business, forecasting, time series prediction, and other applications, are computational systems consisting of artificial neurons called nodes arranged in different layers with interconnecting links. The main interest in neural networks comes from their ability to approximate complex nonlinear functions. Among the available wide range of neural networks, most research is concentrated around feed forward neural networks called Multi-layer perceptrons (MLPs). One of the

important components of an artificial neural network (ANN) is the activation function. This work discusses properties of activation functions in multilayer neural networks applied to breast cancer stage classification. There are a number of common activation functions in use with ANNs. The main objective in this work is to compare and analyze the performance of MLPs which has back-propagation algorithm using various activation functions for the neurons of hidden and output layers to evaluate their performance on the stage classification of breast cancer data.

Survival analysis can be considered a classification problem in which the application of machine-learning methods is appropriate. By establishing meaningful intervals of time according to a particular situation, survival analysis can easily be seen as a classification problem. Survival analysis methods deals with waiting time, i.e. time till occurrence of an event. Commonly used method to classify this sort of data is logistic regression. Sometimes, the underlying assumptions of the model are not true. In model building, choosing an appropriate model depends on complexity and the characteristics of the data that affect the appropriateness of the model. Two such strategies, which are used nowadays frequently, are artificial neural network (ANN) and decision trees (DT), which needs a minimal assumption. DT and ANNs are widely used methodological tools based on nonlinear models. They provide a better prediction and classification results than the traditional methodologies such as logistic regression. This study aimed to compare predictions of the ANN, DT and logistic models by breast cancer survival. In this work our goal is to design models using both artificial neural networks and logistic regression that can precisely predict the output (survival) of breast cancer patients. Finally we compare the performances of these models using receiver operating characteristic (ROC) analysis.

CHAPTER ONE

Introduction

1.1 Cancer

In modern medicine, the term tumor means a neoplasm (from Ancient Greek νεο- neo- "new" and πλάσμα plasma “formation, creation” in field medicine) is an abnormal mass of tissue as a result of uncontrolled growth or division of cells. Some neoplasms do not cause a lump or form an additional tissue. They are called benign. Cancer is a malignant neoplasm or malignant tumor. This malignant neoplasm or tumor is the largest cause for death in United States Cancer. Cancer is not a new disease from the present generation, it has been documented and recorded on a papyrus from ancient Egypt, in 1500 B.C. This oldest document has details that were recorded on a papyrus, documenting 8 cases of tumors occurring on the breast. Further descriptions can be found in ancient writings of Chinese and Arabic literature.

As mentioned earlier, cancer is a condition of abnormal and rapid cell destruction inside the tissues making a mass of extra tissues which is known as tumor. The cancer disease is majorly classified into two types based on the tissue or tumor growth. Benign and malignant. Unlike benign tumors which are assumed not harmful, malignant tumors are formed by jumping of cancer cells to other parts of the body. Scientists have stated the reason behind formation of such condition is due to adhesion property of the cancer causing cells which is stated as the metastasis. The major types of cancers are breast cancer (in women), leukemia (in children),

prostate cancer (in men) and colon cancer. Our present dissertation deals with the subject of breast cancer in women with condition of malignancy. Table 1.1 below gives a brief statistics of estimated deaths of different types of cancers observed in women during 2013 (Source: American Cancer Society).

Table 1.1: Summary of major cancers in women

Different types of cancers in women	Percentages
Lung & Bronchus	72,220 (26%)
Breast	39,620 (14%)
Colon & Rectum	24,530 (9%)
Pancreas	18,980 (7%)
Ovary	14,030 (5%)
Non-Hodgkin Lymphoma	8,430 (3%)
Leukemia	10,060 (4%)
Uterine corpus	8,190 (3%)
Liver & intrahepatic bile duct	6,780 (2%)
Brain/Other nervous systems	6,150 (2%)
All sites	273,430 (100%)

1.2 Breast Cancer

Breast has been considered as a symbol of femininity, fertility and beauty. Breast disease has been known to mankind since old times. Due to the unmistakable side effects particularly at later stages, the bumps that advance into tumors have been recorded by doctors promptly in time. Unlike other inside malignancies, bosom bumps have a tendency to show themselves as noticeable tumors.

Breast cancer is the most common effecting disease in women and second most cause of death for women in United States. It is the cancer that starts in the tissues of the breast with uncontrolled multiplicity affects other parts of the body causing death. There are certain cases of breast cancer observed in men, but it accounts for less than 0.05% of all the cases diagnosed. Breast cancer is classified into two main types:

- Ductal carcinoma: starts in the tubes (ducts) that move milk from the breast to the nipple. Most of the cases fall under this breast cancer.
- Lobular carcinoma: starts in parts of the breast, called lobules that produce milk.

In very rare cases, breast cancer can start in other areas of the breast. According to American Cancer Society (ACS), even at the age of 85 one in eight women are diagnosed with breast cancer. In 2013, an estimated 232,340 new cases of invasive breast cancer are expected to be diagnosed among the women, and about 2240 new cases are expected in men. In addition to this facts, 64,640 new cases of the in situ breast cancer are expected in the women; of which 85% approximately fall into category of ductal carcinoma. One good thing about breast cancer is that it can be treated if it is detected in early stages. The most common outward signs of detection are formation of lumps, or nipple tenderness or thickening of area near the breasts or a dimple in the breast. Less commonly observed signs include breast swelling and enlarged underarm area. The important risk factors include gender, age, family history, early menarche, late menopause, physical inactivity, alcohol consumption, among many others. Other clinical factors for increase in risk are high bone mineral density, biopsy confirmed hyperplasia, high dose radiation to the chest, long menstrual history etc.

1.3 Survival Analysis

Survival analysis today is implemented in almost all fields of sciences. An analysis which is performed to determine the probability of occurrence of the events associated with death or failure after treatment to the subjects is termed as survival analysis. This classification is applicable with help of machine- learning methods that evolve categorical results with predetermined time intervals. Survival analysis of breast cancer has acquired good importance for cancer detection in early stages taking into consideration risk factors. Different kinds of survival studies in present day include clinical trials, prospective cohort studies, retrospective cohort studies and retrospective correlative studies. Survival analysis deals with time to event modeling data with censoring. Censoring is mechanism of identification of the data values which do not follow up until end of the experiment. In many cases data considered for survival analysis are right censored which implies that the concerned subjects leaves the study before the event has occurred or study ends before the event has occurred. The primary interest is to investigate the time to event or the survival probability. The statistical methods employed in study of survival and hazard probability can be performed parametrically, semi-parametrically and non-parametrically based on the nature of the data.

1.3.1 Non-Parametric, Parametric and Semi-parametric Analyses

Non-parametric survival analysis is used to analyze the data avoiding assumptions for the underlying distributions. This kind of analysis restricts the data from occurrence of potential errors. One of the commonly used non-parametric estimator is Kaplan-Meier estimator also called as product limit estimator. The plots of product limit estimator is a graph with declining steps. At times censoring data predicts more accurate results with product limit estimator.

Parametric survival analysis assumes functional form of probability distribution for the variables that provides the influence of explanatory variables on survival time. The strength of this analysis is the estimation is relatively easy and survival curves are smoother as they draw information from whole data. This parametric analysis is carried out using two different approaches which are regression parametric models (Accelerated Failure Time models) and Proportional Hazard (PH models). The name 'accelerated life' is extracted from the industrial applications where the items are subjected to worse conditions than the item usually encounter in real life, so that the experiment is completed in short period of time. Acceleration Failure models are usually applied to the log of the survival time. Different AFT models are generated by assuming different distributions to error term of expression. Estimation of such models using the maximum likelihood is computed for the censored data.

The intermediate model between above two analyses is semi-parametric survival analysis or Cox-regression analysis. It overcomes the disadvantage of the non-parametric analysis of comparing the survival functions for limited number of groups. Cox regression models or PH models are used for the survival time estimation making assumptions to hazard function in the formula. Distribution for the baseline hazard are assumed to follow exponential, Weibull, log-normal, log-logistic or generalized gamma. Even though cox models have driven statistical innovations in past decades, there is more to come in future.

1.4 Logistic Regression

Logistic regression is mostly used to predict a categorical (usually dichotomous) variable from a given set of independent variables. If all the independent variables are continuous, we usually employ discriminant analysis for modeling the data. In case if all or few independent

variables are categorical, logistic regression analysis is the best choice. Also on the other hand, logistic regression makes no assumptions about the distributions of the independent variables. One of the most commonly used tools of medical and clinical applied statistics and discrete data analysis is logistic regression. It is put forward around 1940's against the Fisher's 1936 classification method and considered as center part of many research studies. Logistic regression also finds applications in the fields of engineering, opinion polls, marketing etc.

In logistic regression, the predicted dependent variable is a function of the probability that a particular subject will be in one of the categories (two categories in case of dichotomous dependent variable). In other words, logistic regression is used to predict the probability that the 'event of interest' will occur as a function of one (or more) discrete/continuous and/or dichotomous independent variables (either 0 or 1). For example identifying the relationship between a binary outcome (dependent) variable such as presence or absence of disease when we are given with predictor (explanatory or independent) variables such as patient demographics or imaging findings. The important difference between what is being estimated by a logistic regression model and that estimated by a linear model is that linear regression attempts to predict the value of the dependent variable as a linear function of one or more independent variables. Whereas logistic regression attempts to predict the probability that a unit under analysis will acquire the event of interest as a function of one or more independent variables.

1.5 Artificial Neural Networks

The implementation of artificial neural networks (ANNs) in the field of survivability is suggested to address the limitations of traditional regression methods. ANNs are algorithms which are patterned after the structure of human brain. They possess series of mathematical

equations and terms to simulate the biological process such as learning and memory. Neural networks offers the ability to detect the complex nonlinear relationships between dependent and independent variables. ANNs find applications in the fields of social sciences, clinical studies,

Financial models, altitudes in educational sciences, social mobility, travel behavior, social capital among many others. A basic neural network consists of input, hidden and output layers. The interconnected nodes in different layers possess weights which are adjusted to find the most reliable outcomes by a process termed as learning or training. The most commonly used neural network is multilayer perceptron which consists of one input, one output and one or more hidden layers. The principle of MLP is to reduce the discrepancy between the real and predicted outcomes by propagating discrepancy in backward direction. The merits of trained ANNs is the capability to elevate the information present in the hidden layers without the effect of constraints on the data representation. Limitations of ANNs include its black box nature, greater computation burden, and proneness to over fitting etc. Due to its effective analysis of more complex data, ANNs are used to analyze non-linear covariates, time dependent covariates and versatility among high order covariates. Comparing to traditional regression models ANNs have provided better results concerning to the cancer research.

1.5.1 ANN and Statistics

The artificial neural network (ANNs) and literature in statistics discusses almost same concepts but usually with different terminology. Sometimes the same term in these both literatures may have a different meaning. Below in the Table 1.2 we have mentioned few of such terms used in both the cases.

1.6 Linking ANN, Logistic Regression and Survival analysis

Survival analysis methods deals with waiting time, i.e. time till occurrence of an event. Commonly used method to classify this sort of data is logistic regression. However, sometimes the underlying assumptions of the model may not be true. In model building, choosing an appropriate model depends on complexity and the characteristics of the data that effect the appropriateness of the model. One strategy, which is used nowadays frequently, is artificial neural network (ANN) model which needs a minimal or no assumptions. My current research is aimed to compare survival models and predictions of the ANN models for stage classification, survival and logistic modeling for breast cancer survival.

Table 1.2: ANN and Statistical jargon

Neural networks	Statistics
Architecture	Model
Inputs	Independent (predictor) variable
Outputs	Dependent (outcome) variable, predicted value
Connection weights	Regression coefficients
Bias weight	Intercept parameter
Error	Residuals
Supervised learning	Regression, discriminant analysis
Unsupervised learning	PCA, Data reduction, Clustering
Training set	Sample data
Testing set	Hold-out data
Learning, training	Parameter estimation, fitting
Training case, pattern	Observation
Cross-entropy	Maximum likelihood estimation
Classification	Discriminant analysis
Activation function	Inverse link function in GLIM
Epoch	Iteration

CHAPTER TWO

Parametric Analysis of Breast Cancer Tumor Sizes

2.1 Introduction

Any cancer that grows in our body is always dangerous. If it exists one must try to locate and get it out of our body immediately. Breast cancer is a signature disease of Western populations. Breast cancer is a cancer that starts in the tissues of the breast. There are two main types of breast cancer. Ductal carcinoma starts in the tubes (ducts) that move milk from the breast to the nipple. Most breast cancers are of this type. Lobular carcinoma starts in parts of the breast, called lobules that produce milk (1 –3). In very rare cases, breast cancer can start in other areas of the breast. The three most important things that we can do to find a growth in the breast that may become malignant are: regularly scheduled mammograms, annual clinical breast exams with your health practitioner, and monthly breast self-examination (4).

2.2 Facts and Numbers

Cancer is a major cause of morbidity in the United States, with a total of 1.34 million cases reported during 2005 from 49 of the 50 states (5). According to American Cancer Society (ACS), about 1 in 8 women in the United States (12%) will develop invasive breast cancer over the course of her lifetime (6). In 2016, an estimated 246,660 new cases of invasive breast cancer (includes new cases of primary breast cancer among survivors, but not recurrence of original breast cancer among survivors) are expected to be diagnosed in women in the U.S.,

along with 61,000 new cases of non-invasive (in situ) breast cancer and an estimated deaths due to breast cancer would be around 40,450 (6, 7).

About 2,600 new cases of invasive breast cancer were expected to be diagnosed in men in 2016. Less than 1% of all new breast cancer cases occur in men. For women in the U.S., breast cancer death rates are higher than those for any other cancer, besides lung cancer. Also besides skin cancer, breast cancer is the most commonly diagnosed cancer among U.S. women. More than 1 in 4 cancers in women (about 28%) are effected with breast cancer.

2.3 Questions of Interest

Q1: What is the probability distribution function (PDF) that best characterizes the behavior of malignant tumors for Whites, African Americans and other races?

Q2: Is there any statistical difference between mean tumor sizes between the three races (Whites, African Americans and Others) in the study?

Q3: Is there any statistical difference between mean tumor sizes of any two races?

Q4: If a lady feels a tumor while self-examining, what is the confidence interval estimation for the average tumor size based on her race?

2.4 Data Description

The Surveillance, Epidemiology, and End Results (SEER)-Medicare database links data from the National Cancer Institute's SEER cancer registry program with claims data from Medicare, the federally funded insurance program for the US elderly. These data are made available to investigators and have been used extensively in research (details at http://healthservices.cancer.gov/seermedicare/). This resource is valuable for conducting research on cancers. (8-14)

SEER is a National Cancer Institute-funded program collecting data on cancer incidence and survival from US cancer registries (http://www.seer.cancer.gov). SEER began in 1973 with 9 state and metropolitan area cancer registries. Successive expansions in 1992 and 2001 led to the inclusion in SEER of 17 cancer registries that presently cover approximately 26%of the US population. In total, 146 million person-years are covered during 1973–2007, with 3.1 million incident cancers on the basis of a positive or negative test. The US National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program began collecting the data for many cancers in almost 17 registries.

We obtained breast cancer incidence data from the US National Cancer Institute's SEER program. We used patient and population data from the SEER 9 Registries Database (15, 16) the information that we have used in this present study is obtained from SEER database registry. This data source SEER (16) (Surveillance Epidemiology and End Results), which is a unique, reliable and essential resource for investigating the different aspects of cancer. The SEER database combines patient-level information on cancer site, tumor pathology, stage, and cause of death (3, 4).

In this work, we preprocessed the SEER data (period of 1992-2008 with all records named in breast.txt) for breast cancer to remove redundancies and missing information. The resulting data set had 47,167 malignant tumor records, which then pre-classified into three groups of races. "Whites" (37,341; 79.15%), "African American" (4,234; 9%) and "Others" (5,592; 11.85%) are given in Table 2.1.

Table 2.1 Race and age details

Race	N	Percent	Minimum age	Median age	Maximum age
1	37341	79.17	21	62	102
2	4234	8.98	22	57	102
3	5592	11.86	21	53	99

In this work, demographic information included age, race, and marital status. Tumor characteristics like tumor size (1mm to 998mm), stage of cancer (I, II, III, IV), tumor grade (1, 2, 3, 4, or unknown), and tumor treatment (1, 2, 3, 4) are included.

From Table 2.1, median age at diagnosis in the White women is 62 years (range 21 to 102 years) compared with a median age of 57 years in the African American women (range 22 to 102 years) and a median age of 53 years in the Other races women (range 21 to 99 years). There are 62.15% survival and 37.35% of not survived patients in our data (Table 2.2) and from Table 2.3, majority of patients (about 92%) are diagnosed when they are in stages 1 and 2 and very few (about 8%) of them are diagnosed in advanced stage of breast cancer.

Table 2.2 Survival status details

Status	**Frequency**	**Percent**	**Cumulative percent**
Dead (0)	17853	37.85	37.85
Survived (1)	29314	62.15	100.00

Table 2.3 Breast cancer stage wise details

Stage	**Frequency**	**Percent**	**Cumulative percent**
1	23345	49.49	49.49
2	20017	42.44	91.93
3	2600	5.51	97.45
4	1205	2.55	100.00

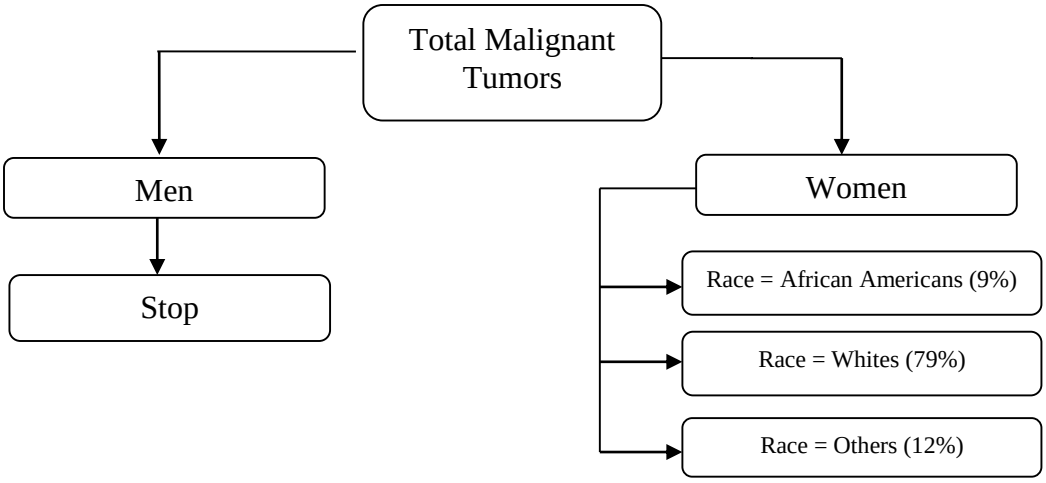


Figure 2.1 Race wise tumor classification chart

2.5 Parametric Analysis of tumor size

Most clinical research involves the collection of some form of quantitative data. The purpose of collecting data is to obtain information that will allow one to infer or draw conclusions about the specific characteristics of a certain large group of subjects or events based on the observation of a few (17 - 20).To select the proper statistical test it is important to know how the data are distributed. The word parametric, or parameter, relates to the nature of data, i.e., the assumptions about particular data. The primary assumptions are that the data points are randomly drawn, that the population is normally distributed and that there is homogeneity among variances. Parametric tests are more stringent than nonparametric tests, and the results tend to be more powerful.

In our work we performed parametric analysis to determine the best fitted distribution that characterizes the behavior of tumor size for each race by setting the hypothesis as follows:

$$H_0: The\ tumor\ size\ data\ followed\ a\ specific\ parametric\ model$$
$$H_1: The\ tumor\ size\ data\ did\ not\ follow\ a\ specific\ parametric\ model$$

After performing many trials, from the class of many parametric distributions, based on the results of minimum Anderson-Darling value, we identified that Inverse Gaussian distribution as the best probabilistic distribution function that characterizes the behavior of the malignant tumors for all the three races considered in this study.

2.5.1 Inverse Gaussian distribution

Over a century, family of Inverse Gaussian distributions had attracted the attention of many researchers in many fields (21). When the data possess some extreme values in it, we need a distribution that can take all the values into consideration, one such is Inverse Gaussian distribution. This is also known as Inverse normal distribution or Wald distribution. Inverse Gaussian distribution is 2-parameter family of continuous probability functions with support on $(0, \infty)$. This distribution is derived while observing the Brownian motion i.e., random movements of atoms and molecules by Schrodinger in 1915 (23).

The Hazard rate function of Inverse Gaussian distribution is uni-modal which increases from zero to its maximum value and decreases asymptotically to a constant. The most differentiating fact is extreme values of outcomes can occur with almost all outcomes being small. It is a right-skewed distribution with long tail. For these reasons Inverse Gaussian distribution is often used in reliability and survival analysis. Various insurance problems and stock markets follow this distribution (22).

The distribution is described by two parameters. Mean or location ($\mu > 0$) and precision or shape ($\lambda > 0$). Let us suppose $x_1, x_2, x_3 \ldots x_n$ be n independent and random variables. If x_i follows the inverse Gaussian distribution, then probability density function of $x_i \sim IG\ (\mu, \lambda)$ is

$$f(x, \theta) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, x \geq 0;\ \theta = (\mu, \lambda)^T$$

The expected value is given by mean μ and variance is equal to$\frac{\mu^3}{\lambda}$. The cumulative distribution function is given by

$$F(y) = \phi(y) + \exp\left(\frac{2\lambda}{\mu}\right)\left(-\sqrt{\frac{4\lambda}{\mu} + y^2}\right);\ -\infty < y < \infty$$

Where□ □is the standard normal distribution function. Clearly, as$\frac{\lambda}{\mu} \to \infty,\ F(y) \to \phi(y)$. The confidence interval for true mean of this distribution is given by $\hat{\mu} \pm z_{\alpha/2}(n\lambda)^{-1/2}\hat{\mu}^{3/2}$

2.5.2 PDF for White women

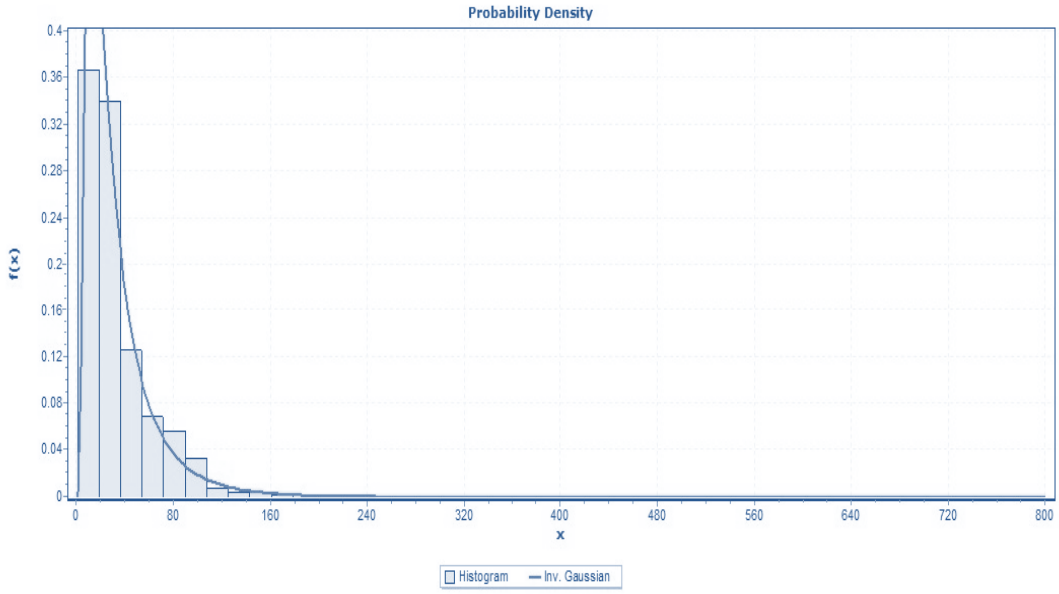


Figure 2.2 PDF for white women: Inverse Gaussian Distribution

The fitted PDF and CDF of tumor sizes for white race women is

$$f(x,\theta) = \left(\frac{43.93}{2\pi x^3}\right)^{\frac{1}{2}} exp\left\{-\frac{43.93(x-32.76)^2}{2(32.76)^2 x}\right\}, x \geq 0;$$

$$F(y) = \phi(y) + \exp(2.682)\left(-\sqrt{5.36 + y^2}\right)$$

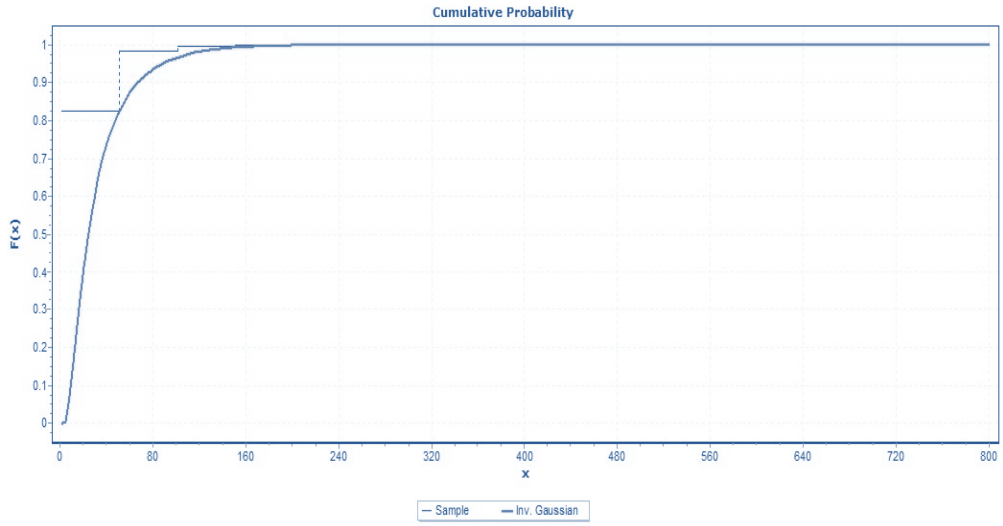


Figure 2.3 Inverse Gaussian CDF for White Women

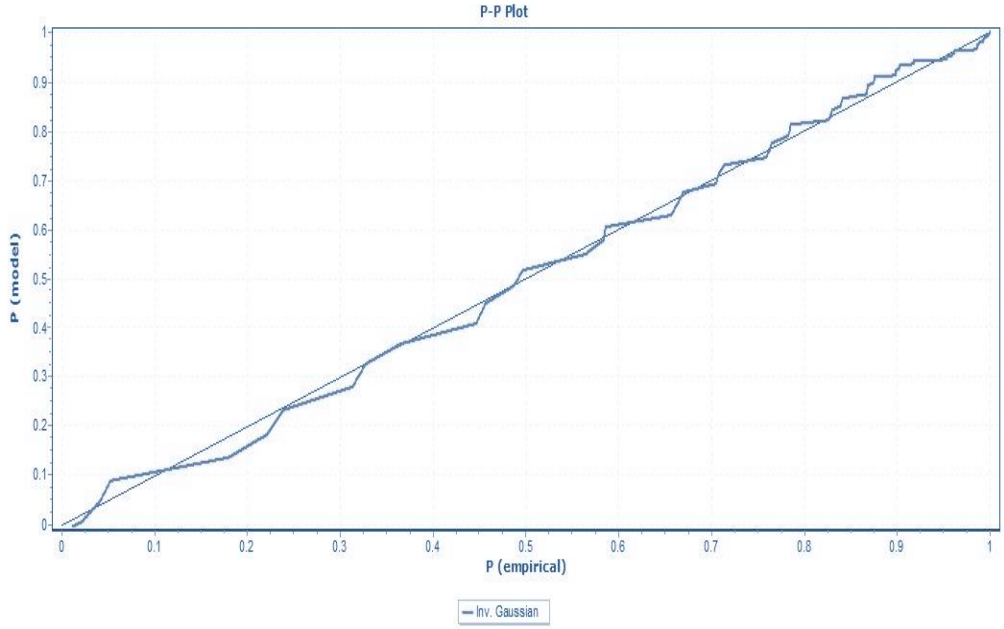


Figure 2.4 Inverse Gaussian PP Plot for White Women

Figure 2.2 is the fitted Inverse Gaussian PDF with estimated shape and location parameters as 43.933 and 32.756 respectively. From Figure 2.3 the CDF graph explains how well the distribution fit to data and the PP plot in Figure 2.4 is approximately linear and confirms about the fitted distribution.

2.5.3 PDF for African American women

Figure 2.5 below is the fitted Inverse Gaussian PDF for AA women with estimated shape and location parameters as 66.614 and 39.611 respectively. From Figure 2.6 the CDF graph explains how well the distribution fits to data and the PP plot in Figure2.7 is approximately linear and confirms about the fitted distribution.

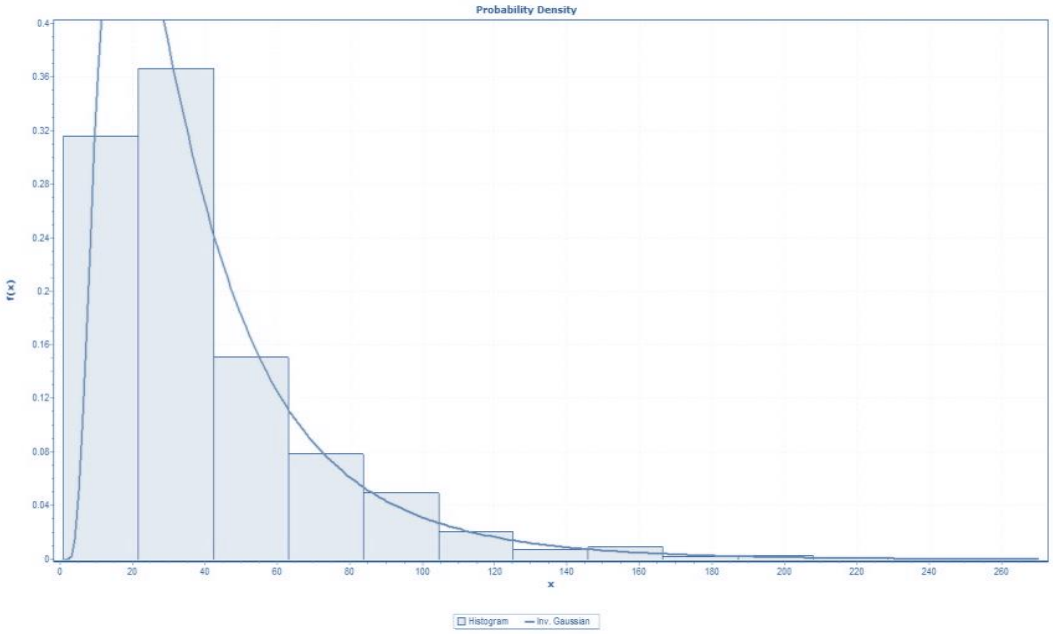


Figure 2.5 PDF for African American Women: Inverse Gaussian distribution

The fitted PDF and CDF of tumor sizes for African American race women is

$$f(x,\theta) = \left(\frac{66.61}{2\pi x^3}\right)^{\frac{1}{2}} exp\left\{-\frac{66.61(x-39.61)^2}{2(39.61)^2 x}\right\}, x \geq 0;$$

$$F(y) = \phi(y) + \exp(3.363)\left(-\sqrt{6.73 + y^2}\right)$$

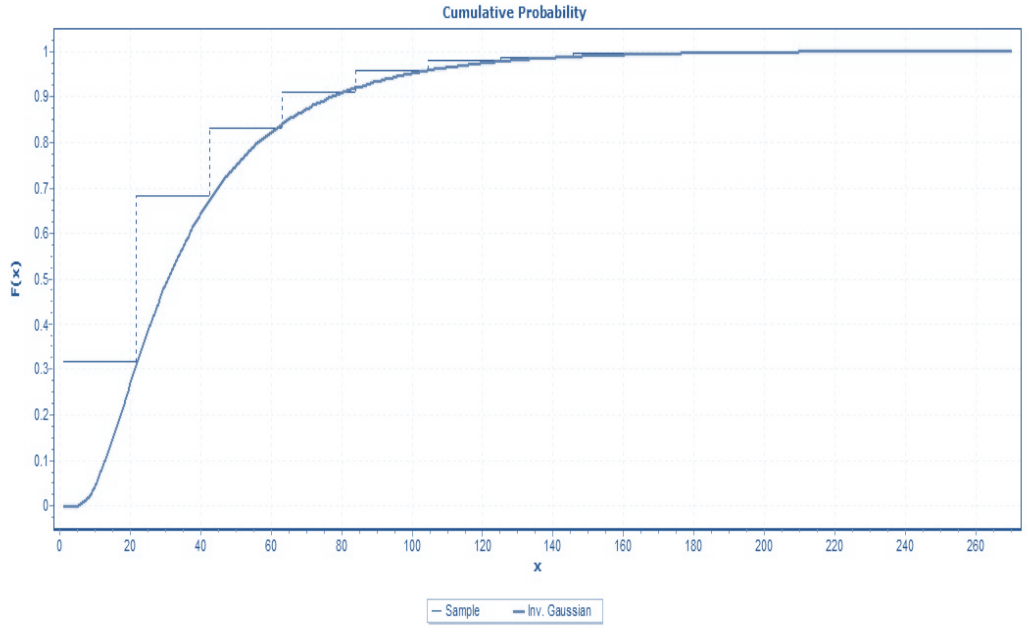


Figure 2.6 Inverse Gaussian CDF for African American Women

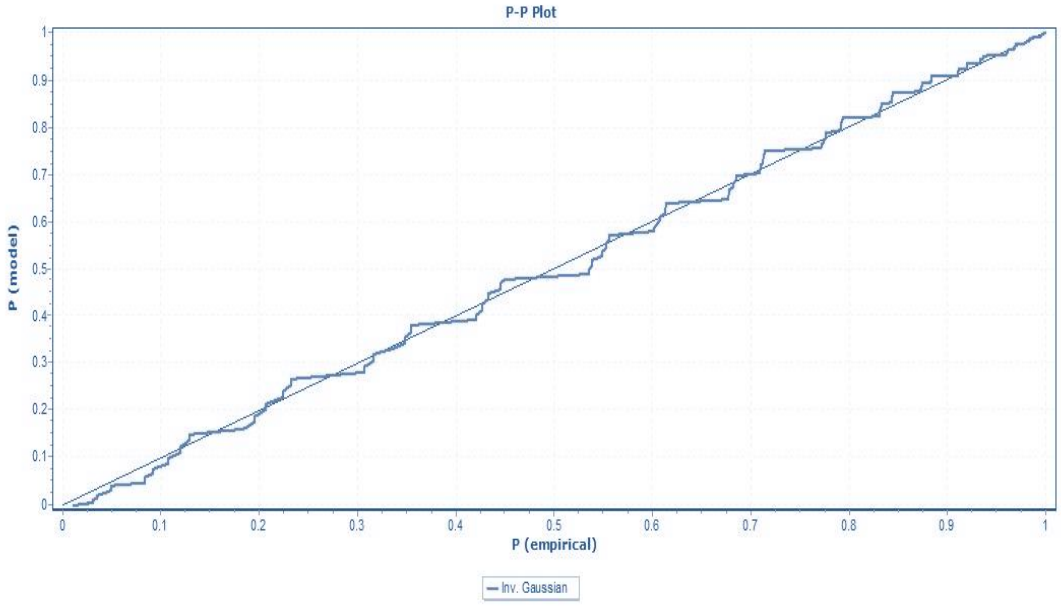


Figure 2.7 Inverse Gaussian PP Plot for African American Women

2.5.4 PDF for Other Races

Figure 2.8 below is the fitted Inverse Gaussian PDF for other race women with estimated shape and location parameters as 55.703 and 36.846 respectively. From Figure 2.9 the CDF graph explains how well the distribution fit to data. The fitted PDF and CDF of tumor sizes for other race women is

$$f(x,\theta) = \left(\frac{55.70}{2\pi x^3}\right)^{\frac{1}{2}} exp\left\{-\frac{55.70(x-36.85)^2}{2(36.85)^2 x}\right\}, x \geq 0;$$

$$F(y) = \phi(y) + \exp(3.02)\left(-\sqrt{6.05+y^2}\right)$$

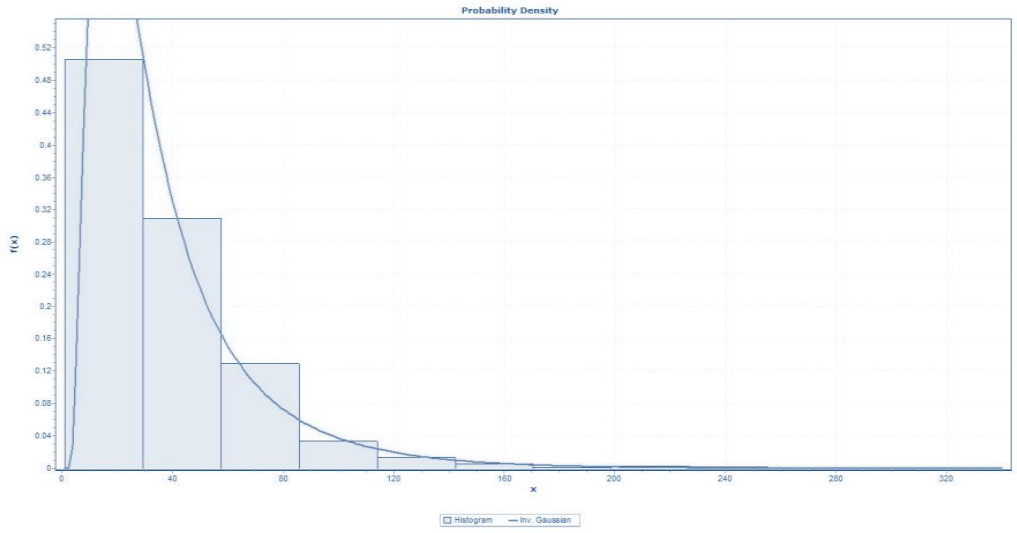


Figure 2.8 PDF for Other races: Inverse Gaussian Distribution

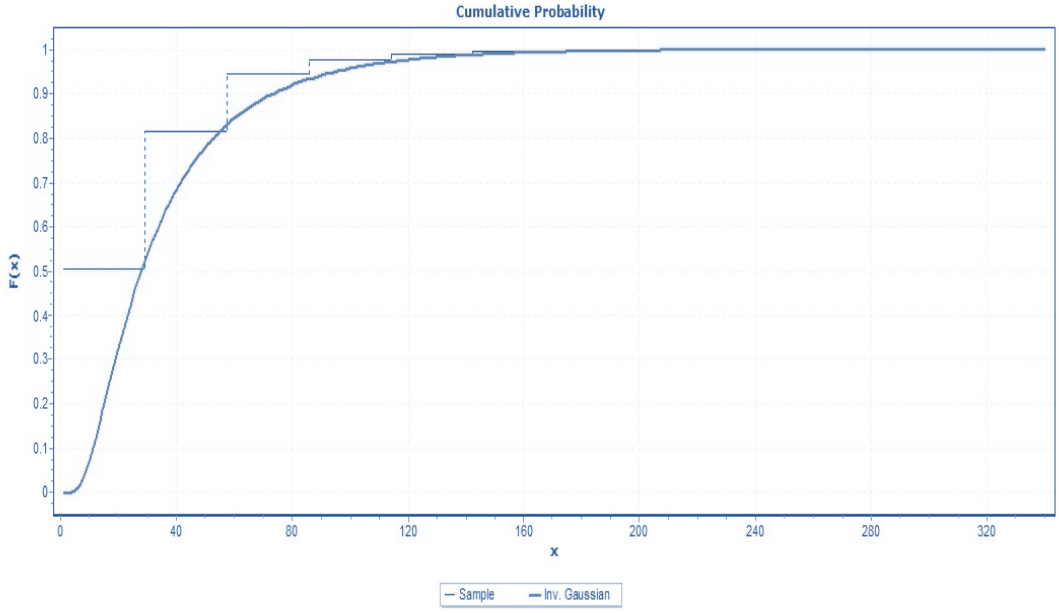


Figure 2.9 Inverse Gaussian CDF for Other race women

2.5.5 Summary of PDF's

Table 2.4 PDF summary for three races

RACE	$\hat{\lambda}$	$\hat{\mu}$
White	43.933	32.756
African American	66.614	39.611
Others	55.703	36.846

Table 2.5 has the race wise details of 95% and 99% confidence interval estimation of true mean tumor size based on Inverse Gaussian distribution. After identifying the distribution functions that best characterizes the probability distribution of malignant tumors for the three races, we proceed to compare the differences of mean tumor sizes for the three races.

Table 2.5 The mean tumor size and confidence intervals of all the three races

Race	$\hat{\mu}$	$\hat{\lambda}$	SD	95% CI for μ	99% CI for μ
1	32.756	43.933	28.284	(32.47, 33.04)	(32.38, 33.13)
2	39.611	66.614	30.545	(38.69, 40.53)	(38.40, 40.82)
3	36.846	55.703	29.967	(36.06, 37.63)	(35.81, 37.88)

2.6 Comparison of mean tumor sizes

Let $\mu_w, \mu_{aa}, and\ \mu_{oth}$ represent mean tumor sizes of whites, African Americans and other races respectively. Our interest is to test the hypothesis whether all the three races have same mean tumor size or otherwise.

$$H_0: \mu_w = \mu_{aa} = \mu_{oth}\ vs.\ H_1:\ At\ least\ one\ of\ them\ is\ not\ equal.$$

By performing a one way ANOVA at 5% level of significance, we obtained the p-value which is very low($p < 0.0001$); leading us to the conclusion that there is significant difference between the average tumor sizes of all the three races. So, we now proceed in pair wise testing of mean tumor sizes for all three races. The Table2.6 below has the details of the results after performing t-test for pair wise testing. Clearly, we conclude that the average tumor size is significantly different for all the three races in this study.

Table 2.6 Pair wise comparison of mean tumor sizes

H_{Null}	$H_{Alternative}$	P-value	Conclusion	95% CI for mean differences
$\mu_w = \mu_{aa}$	μ_w not equals μ_{aa}	0.001	Reject Null	(8.107, 8.659)
$\mu_{aa} = \mu_{oth}$	μ_{aa} not equals μ_{oth}	0.0001	Reject Null	(-10.191, -7.760)
$\mu_w = \mu_{oth}$	μ_w not equals μ_{oth}	0.0002	Reject Null	(-18.547, -16.171)

Previous studies (24 - 26) have shown that breast cancer in these younger women is more aggressive, with higher rate of occurrence and recurrence rates compared with older women. In our study we have the median age of women for all the three races more than 50 years. In Table 2.7, we classified the tumor stage taking age group into consideration. The majority of women are in the ages from 45 to 79. From Table 2.8 and Figure 2.11, African American women are the majority of population in all the age groups who are diagnosed with breast cancer. Table 2.8 gives the age group wise confidence interval for mean tumor size for all the three races. Very interestingly, from Figure 2.10 majority of women in younger ages (20 – 44 years) are identified with stage-2 breast cancer.

Table 2.7 Age group Vs. Stage classification

AGE	**Stage 1**		**Stage 2**		**Stage 3**		**Stage 4**		**All**
	Count	Row%	Count	Row%	Count	Row%	Count	Row%	Total
20-24	7	25.93	17	62.96	3	11.11	0	0.00	27
25-29	67	26.59	152	60.32	25	9.92	8	3.17	252
30-34	239	27.86	481	56.06	94	10.96	44	5.13	858
35-39	700	34.08	1157	56.33	152	7.40	45	2.19	2054
40-44	1456	37.89	1992	51.83	305	7.94	90	2.34	3843
45-49	2153	41.54	2560	49.39	348	6.71	122	2.35	5183
50-54	2561	45.78	2546	45.51	342	6.11	145	2.59	5594
55-59	2634	50.44	2209	42.30	250	4.79	129	2.47	5222
60-64	2723	53.92	1999	39.58	209	4.14	119	2.36	5050
65-69	2909	56.58	1892	36.80	199	3.87	141	2.74	5141
70-74	2997	59.35	1755	34.75	169	3.35	129	2.55	5050
75-79	2496	58.03	1509	35.08	199	4.63	97	2.26	4301
80-84	1526	55.49	993	36.11	143	5.20	88	3.20	2750
85+	877	47.61	755	40.99	162	8.79	48	2.61	1842
All	23345	49.49	20017	42.44	2600	5.51	1205	2.55	47167

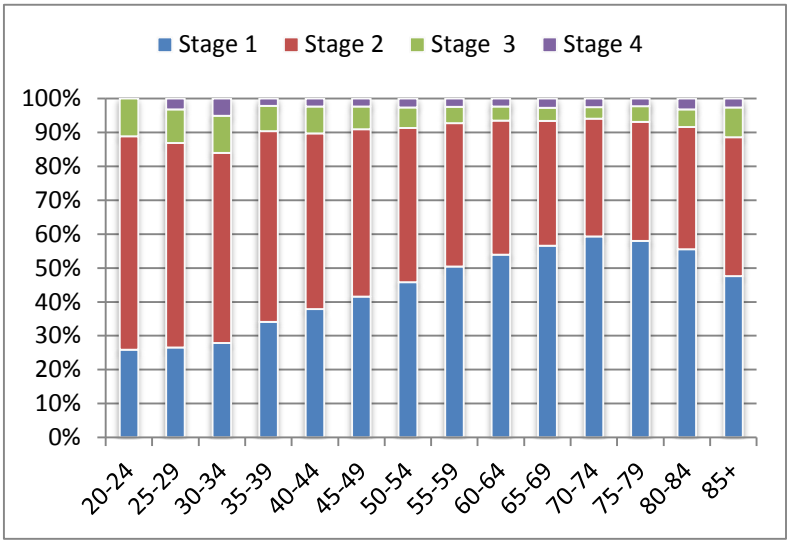


Figure 2.10 Stages vs. age group

Table 2.8 Age group based race wise confidence interval of tumor sizes

Age Group	Race 1				Race 2				Race 3			
	Mean	S.D	C.I (95%)		Mean	S.D	C.I (95%)		Mean	S.D	C.I (95%)	
			L.C.I	U.C.I			L.C.I	U.C.I			L.C.I	U.C.I
20-24	27.12	15.75	19.40	34.84	26.86	13.67	16.73	36.99	18.5	3.32	15.25	21.75
25-29	33.53	74.51	22.70	44.36	56.2	162.8	3.02	109.38	54.2	167.6	-2.14	110.54
30-34	36.24	82.99	29.51	42.97	46.8	124.2	24.94	68.66	28.55	22.25	24.99	32.11
35-39	29.44	60.16	26.36	32.52	28.46	22	25.57	31.35	39.19	114.91	27.37	51.01
40-44	27.79	59.2	25.58	30.00	40.23	110.44	29.35	51.11	26.84	55.2	22.74	30.94
45-49	27.81	70.22	25.57	30.05	35.8	89.67	27.88	43.72	27.11	67.33	22.71	31.51
50-54	25.06	62.78	23.15	26.96	36.92	95.54	28.89	44.95	24.01	38.44	21.44	26.58
55-59	24.44	68.85	22.31	26.57	28.56	64.53	23.03	34.09	23.81	55.75	19.62	28.00
60-64	22.20	59.49	20.36	24.04	26.27	51.07	21.63	30.91	22.21	44.1	18.58	25.84
65-69	23.03	67.36	20.99	25.07	35.11	104.32	25.51	44.71	19.52	16.65	18.05	20.98
70-74	20.41	53.55	18.81	22.00	30.67	78.75	22.47	38.87	29.45	101.49	19.18	39.72
75-79	21.31	48.48	19.75	22.86	34.58	98.68	23.52	45.64	19.84	14.66	18.10	21.58
80-84	22.65	51.02	20.62	24.68	26.81	21.21	23.79	29.83	21.89	25.49	17.49	26.29
85+	29.43	73.19	25.91	32.95	30.42	23.38	26.20	34.64	25.12	16.65	21.10	29.14

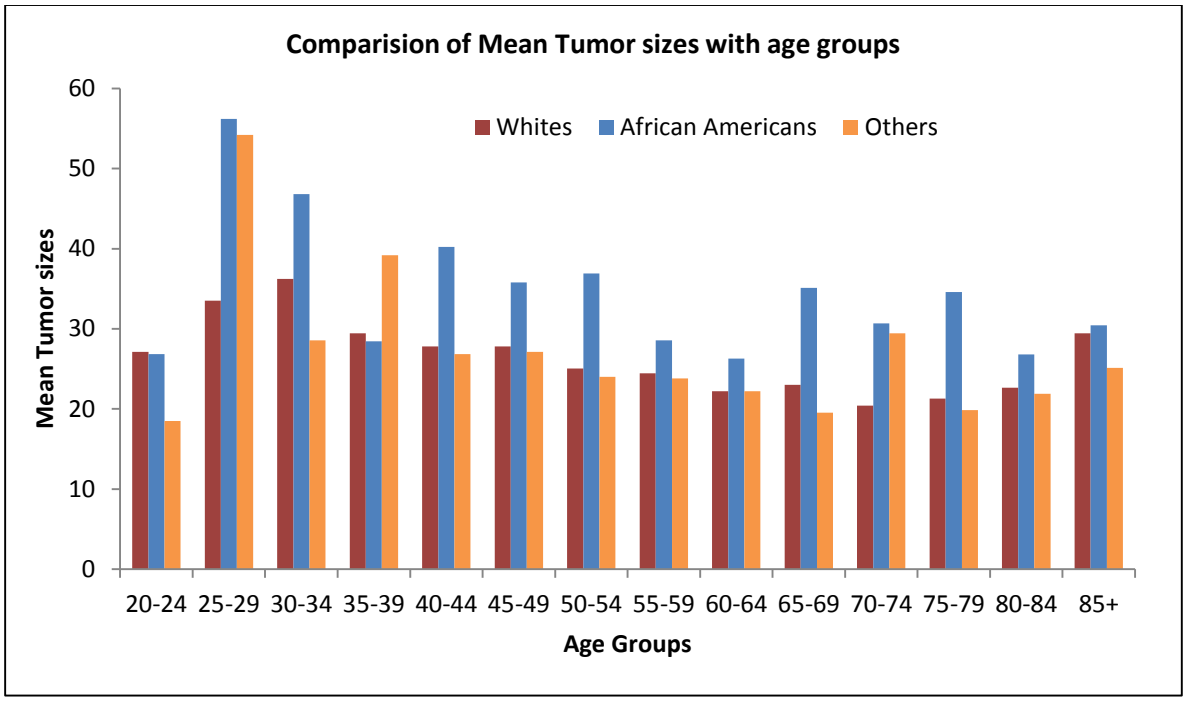


Figure 2.11 Race wise comparison of mean tumor sizes

2.7 Conclusion

The PDF for all the three races is identified as Inverse Gaussian and the details about mean tumor sizes along with 95% and 99% confidence intervals for mean tumor sizes for all the three races were tabulated in Table 2.5. One way ANOVA was performed for comparing mean tumor sizes of three races and at 5% level of significance, we conclude that the average tumor size for all the three races is statistically not the same. Later, we performed pair-wise testing between the races and the results are tabulated in Table 2.6. From these results we conclude that the average tumor sizes are significantly different for all the three races. Also compared with Whites and other race women, African American women have comparatively a greater mean tumor sizes and Whites have the least. This is also supported by the results published in Table 2.7. Finally grouping ages into groups of 5, we also stratified the number of women diagnosed with breast cancer in different stages and Table 2.8 gives the race wise confidence intervals.

CHAPTER THREE

Statistical Analysis on Survival times of Breast Cancer Data

3.1 Introduction

Cancer is a major cause of morbidity in the United States, with a total of 1.34 million cases reported in the year 2005 from 49 of the 50 states (6). Cancer incidence typically rises with age, and a disproportionate fraction of cases occur among the elderly. According to the statistical sources, today in the United States, approximately one in eight women over their lifetime have a risk of developing breast cancer. The statistical methods for survival analysis have been extracted from the biomedical and epidemiologic studies of humans and animals. Basically, survival analysis has its application in data evaluation on the length of time it takes for occurrence of a specific event of interest. The event of interest can be death of person or an animal or any living being or study of termination of particular equipment. One can identify the survival rate with a possibility of data collection related to a particular disease. From the recent data the survival rate of patient with breast cancer is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis (27).

In his book ‘Natural and Political Observations upon the Bill of Mortality’, John Graunt's classified registered deaths by age, period, gender and cause of death, suggested for the first time that death be regarded as an event which deserves systematic study (28,29). Survival data is mainly concerned with time or study analysis of subject or event of interest. This data may also contain subjects which have not experienced its effect over a time or complete study of

analyticity. For instance, some patients may still be alive at the end of a study period. For these subjects, the exact survival times are unknown. This scenario can also be exhibited when the individuals do not follow-up after certain medical attention after a period of study. This would be beyond the practical limits to wait until every subject has died before conducting any analysis which is an intrinsic characteristic of survival data. This pattern of behavior cannot be validated to military and defense officers. Their survival time is usually estimated as the length of survival time at the time of leaving service and becoming the reserve. The officers that are still active at end of the study period are treated as censored observations. Further studies like data collection, evaluation and results related to objective are discussed in following sections.

3.2 Questions of Interest

Q1: Is there a significant difference in the average survival time between the three races?

Q2: Is there a significant difference in the average survival of any two races?

Q3: What is the appropriate probability distribution function (PDF) that best characterizes the survival time of subjects under study for Whites and African Americans and other races?

Q4: What is the behavior of survival functions for all the three races?

Q5: What are hazard and cumulative hazard curves explaining the behavior of the variable of interest?

3.3 Data Description

The information that we have used in this present study is obtained from SEER database registry. This data source SEER (Surveillance Epidemiology and End Results), which is a unique, reliable and essential resource for investigating the different aspects of cancer. The

SEER database combines patient-level information on cancer site, tumor pathology, stage, and cause of death (15, 16). In this work, we preprocessed the SEER data (period of 1992-2008 with all records named in breast.txt) for breast cancer to remove redundancies and missing information. The resulting data set provide 47,167 records, which then pre-classified into two groups of "survived" (29,314; 62.15%) and "not survived" (17,853; 37.85%). The "survived" class is all records that have a duration period value greater than or equal 204 months and the "not survived" class represent the remaining records. In all these cases of breast cancer women analyzed, which included 79.17% White women, 8.98% African American (AA), and 11.86% other races women (American Indian/AK native, Asian/ Pacific Islander). Our primary variable of interest here is the survival time and its probabilistic behavior. The overall description is provided in Table 3.1 and Figure 3.1 provides with the race wise descriptive statistics of the survival time. Table 3.3 and Table 3.4 have the details about race wise and treatment wise survival or otherwise of women considered in our data.

Table 3.1 Descriptive Statistics of survival time in months

Race	Sample Size	Range	Mean	Variance	Median	C.V
Whites	37341	202	100.05	2512.7	98	0.50102
AA	4234	202	89.183	2742.1	84	0.58717
Others	5592	202	101.5	2201.5	97	0.46225

Table 3.2 Survival based classification

Censor	Frequency	Percent	Cumulative %
0 (Dead)	17853	37.85	37.85%
1 (Censored)	29314	62.15	100.00%

Table 3.3 Race wise survival classification

Race	Coded		Censor = 0	Censor = 1	Total
Whites (W)	1	Frequency	14229	23112	37341
		Percent	38.1	61.9	100
African Americans (AA)	2	Frequency	1992	2242	4234
		Percent	47.0	53.0	100
Others (Oth)	3	Frequency	1632	3960	5592
		Percent	29.2	70.8	100

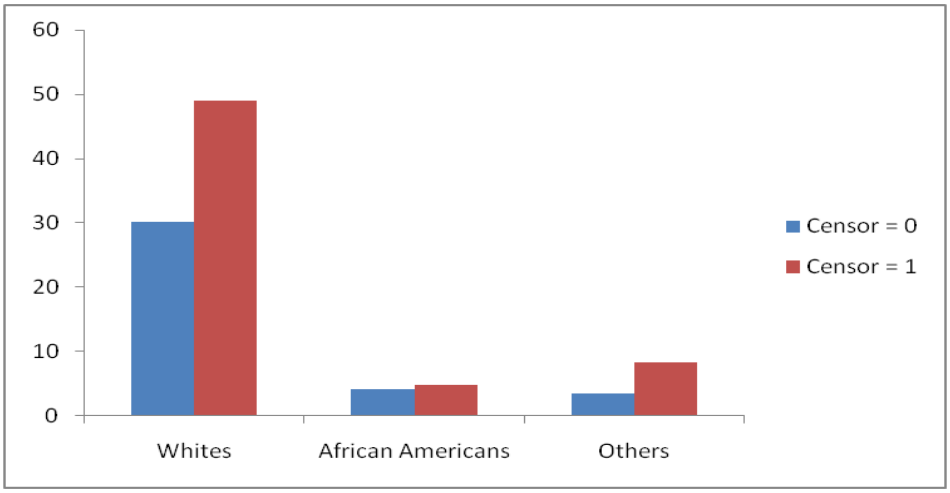


Figure 3.1 Race wise survival classification

Table 3.4 Treatment Classification

Treatment	Coded		Censor = 0	Censor = 1	Total
No treatment	1	Frequency	6053	14656	20709
		Percent	29.2	70.8	100
Radiation	2	Frequency	11116	14489	25605
		Percent	43.4	56.6	100
Radiation & Surgery	3	Frequency	182	33	215
		Percent	84.7	15.3	100
Surgery	4	Frequency	502	136	638
		Percent	78.7	21.3	100

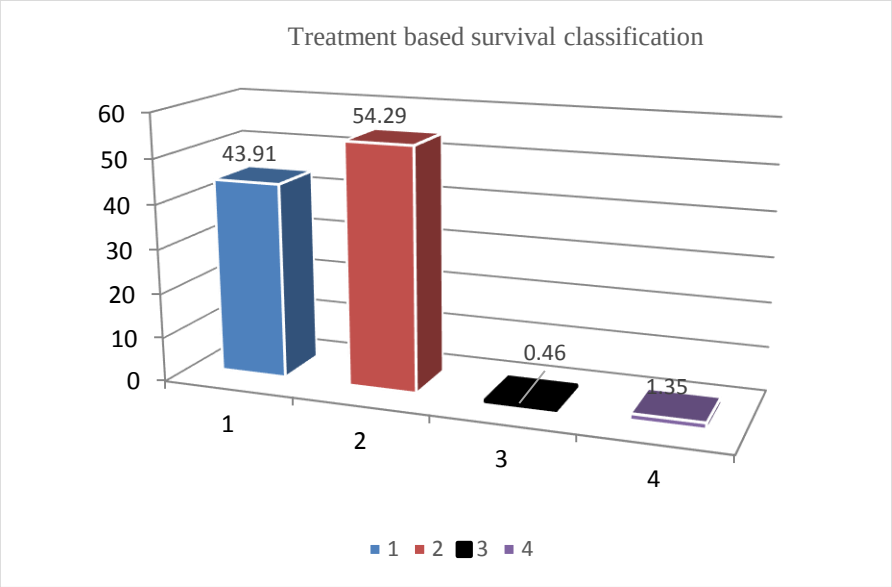


Figure 3.2 Treatment based survival classification

3.4 Comparing Survival times

Kaplan Meier (KM) curve (30, 31) or the product-limit survival plot indicates the unconditional probability that a subject will survive beyond time *t* but do not indicate the proportion of subjects surviving to time *t*. Since all observations are considered alive at beginning of study, the KM survivor function starts at 1 and declines as subjects fail over time. From the Figure 3.3, we can see that the survival probability of an observation lasting beyond time period 100 months is about 0.7 for White race women, 0.58 for African American women and 0.78 for other race women. And the survival probability of a women with breast cancer surviving beyond time 150 months is about 0.56 for White women, 0.48 for African women and 0.64 for other race women.

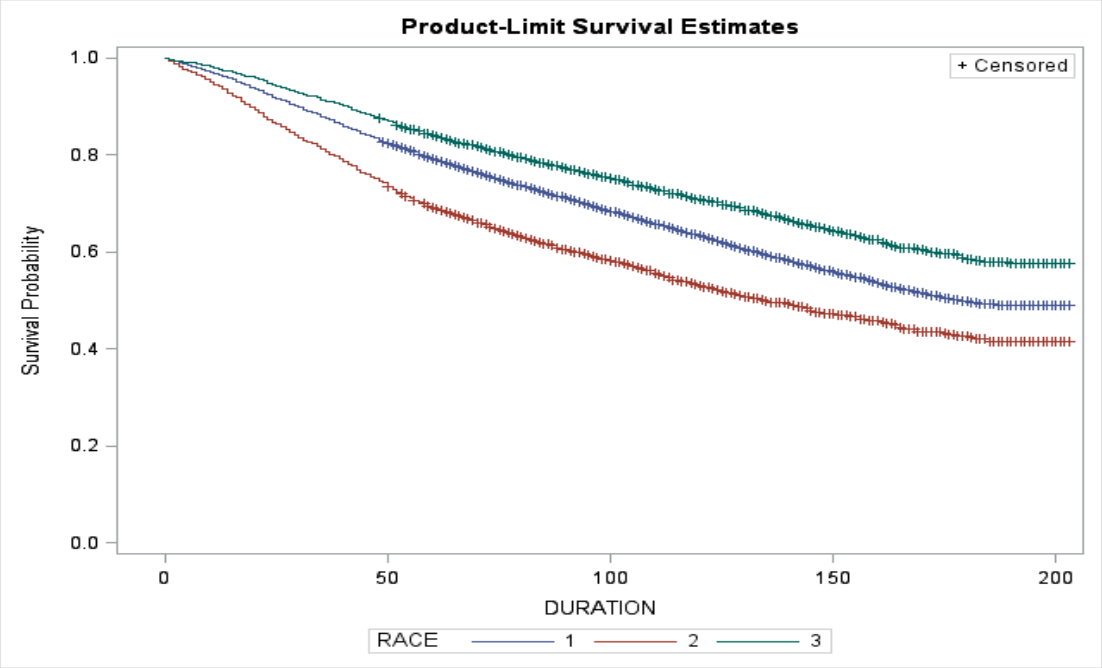


Figure 3.3 Product-Limit survival probability of the three races

Table 3.5 Race Wise Summary Statistics for duration

Summary of the Number of Censored and Uncensored Data Values					
Stratum	**RACE**	**Total**	**Censored**	**Failed**	**Censored (%)**
1	W	37341	23112	14229	38.11
2	AA	4234	2242	1992	47.05
3	Others	5592	3960	1632	29.18
Total		47167	29314	17853	37.85

Table 3.6 Test of equality between three races

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**$Pr >$ Chi-Square**
Log-Rank	346.8230	2	<.0001
Wilcoxon	403.2763	2	<.0001
-2Log(LR)	332.1676	2	<.0001

Results of the comparison of survival curves between the three races are shown in Figure 3.2, Table 3.5 and Table 3.6. Table 3.5 has details about race wise censored data followed by test of equality over the three races in Table 3.6. From Table 3.5, there were a total of 17853 women (38%) who died of breast cancer. There were a total of 29314 women (62%) that were alive at the last assessment period. Also, the log-rank test, which places more weight on larger survival times, is more significant than the Wilcoxon test, which places more weight on early survival times. Clearly, the rank tests for homogeneity in Table 3.6 indicate a significant difference between survival times between all the three the races ($p < 0.0001$ for the log-rank test and $p < 0.0001$ for the Wilcoxon test). From Figure 3.3, other race women live significantly longer than White and African American race women, while African American women comparatively have less survival.

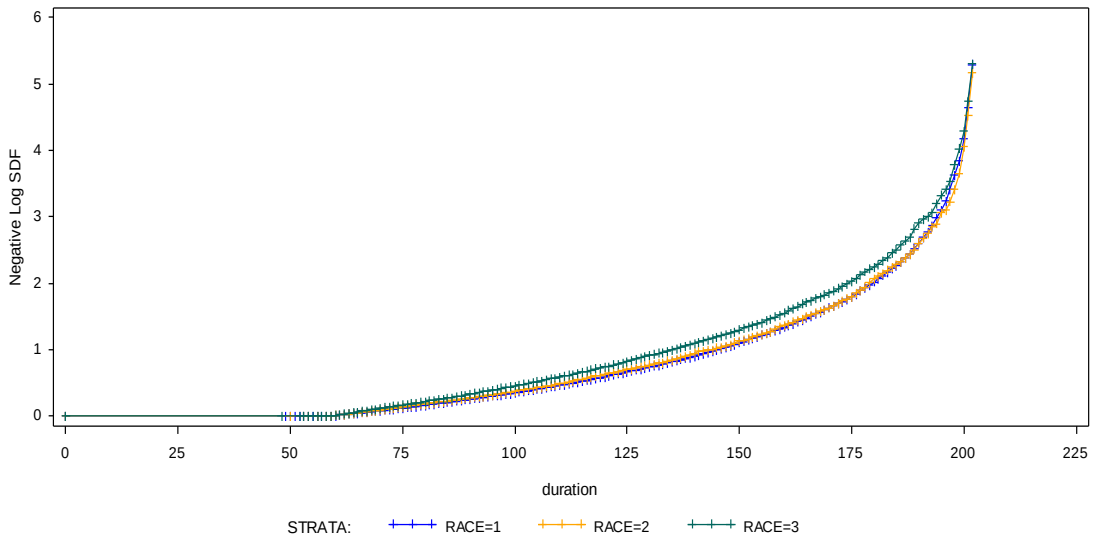


Figure 3.4 Negative Log Survival DF

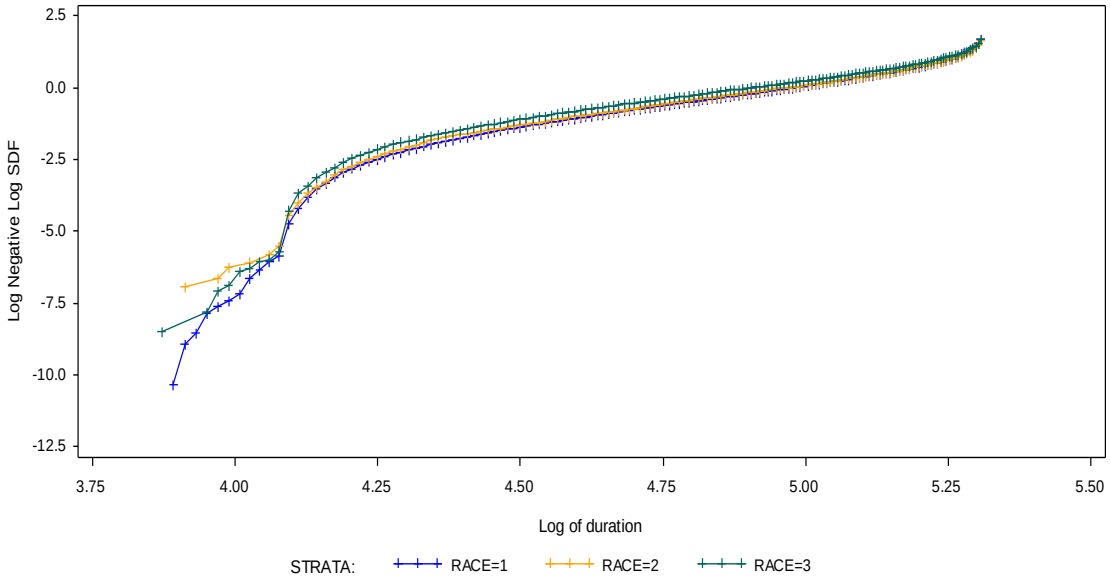


Figure 3.5 Log Negative Log vs. Log Duration Survival DF

A plot of the estimated survivor function against time, a plot of the negative log of the estimated survivor function against time, and a plot of the log of the negative log of the estimated survivor function against log time are given in Figure 3.3, Figure 3.4, and Figure 3.5 respectively. Figure 3.4 and Figure 3.5 provide an empirical check of the appropriateness of the exponential model and the Weibull model, respectively, for the survival data.

If the exponential model is appropriate, the curve in Figure 3.4 should be approximately linear through the origin. Clearly from Figure 3.4 we cannot proceed with exponential model. If the Weibull model is appropriate, the curve in Figure 3.5 should be approximately linear. From Figure 3.5, we can notice a non-linear trend in the data, which stops us to proceed even with Weibull model. Since there is more than one stratum, the Figure 3.5 plot may also be used to check the proportional hazards model assumption. Under this assumption, the log of the negative log of the estimated survivor function curves should be approximately parallel across strata, which in this case fails.

3.5 Parametric Analysis

Probability theory defines distribution by histogram of survival times, given by probability density function (PDF) f (t), cumulative distribution function (CDF) which is the cumulative area under histogram starting from left, given by $F(t) = \int_{-\infty}^{t} f(x)dx$, survivor function $S(t) = 1 - F(t)$, hazard function $h(t) = \frac{f(t)}{S(t)}$ and cumulative hazard function $H(t) = \int_{0}^{t} h(x)dx$.

3.5.1 Probability Density Function

The probability density function (PDF) is also very useful in describing the continuous probability distribution of a random variable. The PDF of a random variable T, denoted$f(t)$, is defined by$f(t) = d\,F(t) / dt$, where $F(t)$ is the cumulative density function (CDF).That is, the pdf is the derivative or slope of the cumulative density function (CDF),$F(t)$. Every continuous random variable has its own density function, the probability $P(a < T < b)$ is the area under the curve between a, b. In this chapter we tried to identify the best fit probability function that characterizes the survival time for all the three races (Whites, AA & Others) separately. We have identified Generalized Extreme Value distribution (GEV) as the best fit for both White and African American races with -0.25296, 81.455, 49.931 and -0.17371, 67.907, 49.663 as estimated shape, location and scale parameters for Whites and African American women respectively. Lognormal is identified as the best fit for the other race women with estimated shape, location and scale parameters as 0.07439, -529.26, and 6.4442. Figure 3.6, Figure 3.7, Figure 3.8 are the PDF's for the three races and Table 3.7 gives the details of the parameter estimates of the fit.

The *Generalized Extreme Value* (GEV) distribution (32 – 34) is a flexible three-parameter model that combines the Gumbel, Fréchet, and Weibull *maximum* extreme value distributions. GEV also has a link to logit functions. GEV has the following analytic form of PDF,

$$f(x) = \begin{cases} \frac{1}{\sigma}\exp\{-(1+kz)^{-1/k}(1+kz)^{-1-(1/k)} & for\ k \neq 0 \\ \frac{1}{\sigma}\exp(-z-\exp(-z)) & for\ k = 0 \end{cases}$$

k, *σ*, *μ* are the shape, scale, and location parameters respectively. The scale must be positive (*σ* >0), the shape and location can take on any real value. The range of definition of the GEV distribution depends on$k$. Specifically, the three cases $k = 0, k > 0, and\ k < 0$ correspond to the Gumbel, Fréchet, and "reversed" Weibull distributions.

The three parameter *lognormal distribution* (34, 35) is based on the Normal distribution. A random variable is log normally distributed if the logarithm of the random variable is normally distributed. With $x > \mu \geq 0$; $-\infty < \sigma < \infty$; $k > 0$, and μ is the location parameter, that defines the point where the support set of the distribution begins; σ is the scale parameter that stretch or shrink the distribution and *k* is the shape parameter that affects the shape of the distribution the probability distribution function of three parameter lognormal distribution function and its corresponding cumulative distribution function (CDF) are given by:

$$f(x) = \frac{1}{(x-\mu)k\sqrt{2\pi}} exp\left\{-\frac{[\ln(x-\mu)-\sigma]^2}{2k^2}\right\}$$

$$F(x) = \Phi\left\{\frac{\ln(x-\mu)-\sigma}{k}\right\}$$

FOR WHITE RACE WOMEN: The fitted GEV distribution that characterizes the breast cancer survival time for White race women is $(x) = \frac{1}{49.931}\exp\{-(1+(-0.253)z)^{-1/(-0.253)}(1+(-0.253)z)^{-1-(-0.253)}\}$; $where\ z = \frac{x-81.455}{49.931}$. The graph of the fitted distribution is given in Figure 3.6.

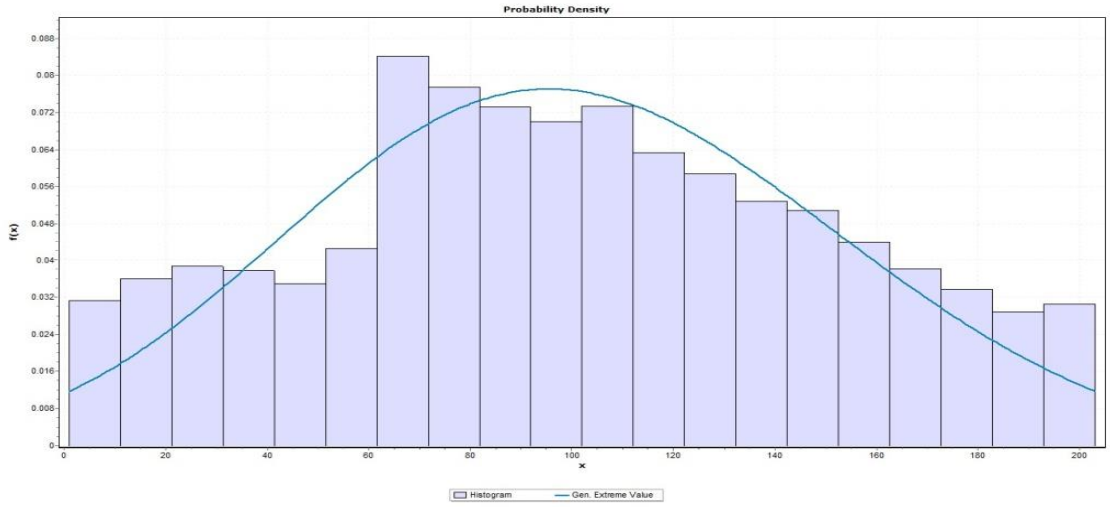


Figure 3.6 PDF of White (GEV distribution)

FOR AA RACE WOMEN: The fitted GEV distribution that characterizes the breast cancer survival time for AA race women is $f(x) = \frac{1}{49.663}\exp\{-(1+(-0.174)z)^{-1/(-0.174)}(1+(-0.174)z)^{-1-(-0.174)}\}$; $where\ z = \frac{x-67.907}{49.663}$. The graph of the fitted distribution is given in Figure 3.7.

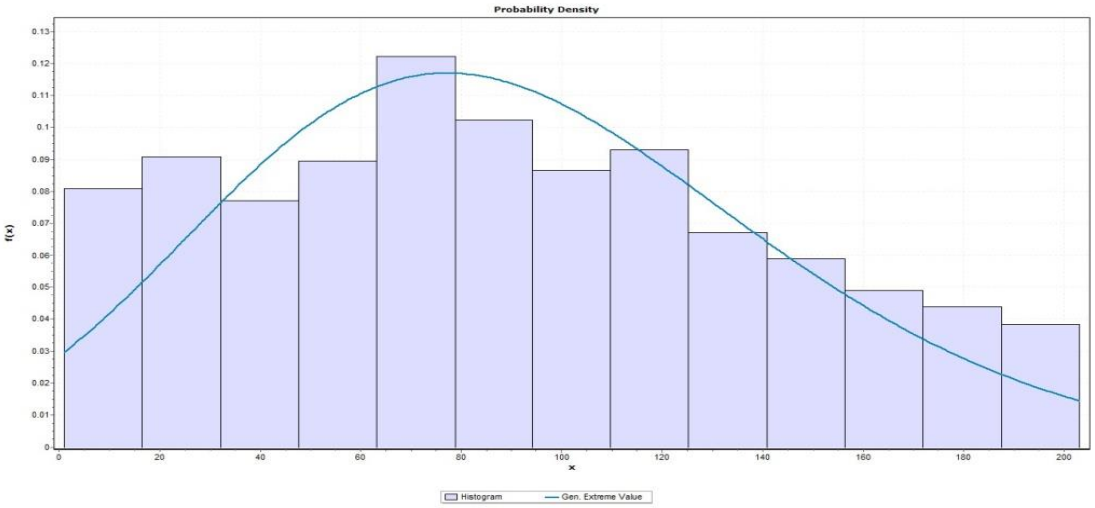


Figure 3.7 PDF of AA race (GEV distribution)

FOR OTHER RACES WOMEN: The fitted lognormal distribution that characterizes the breast cancer survival time for other race women is

$$f(x) = \frac{1}{(0.0744)(x-(-529.26))\sqrt{2\pi}} exp\left\{-\frac{[\ln(x-(-529.26))-6.444]^2}{2(0.0744)^2}\right\}.$$

The PDF graph is given in Figure 3.8.

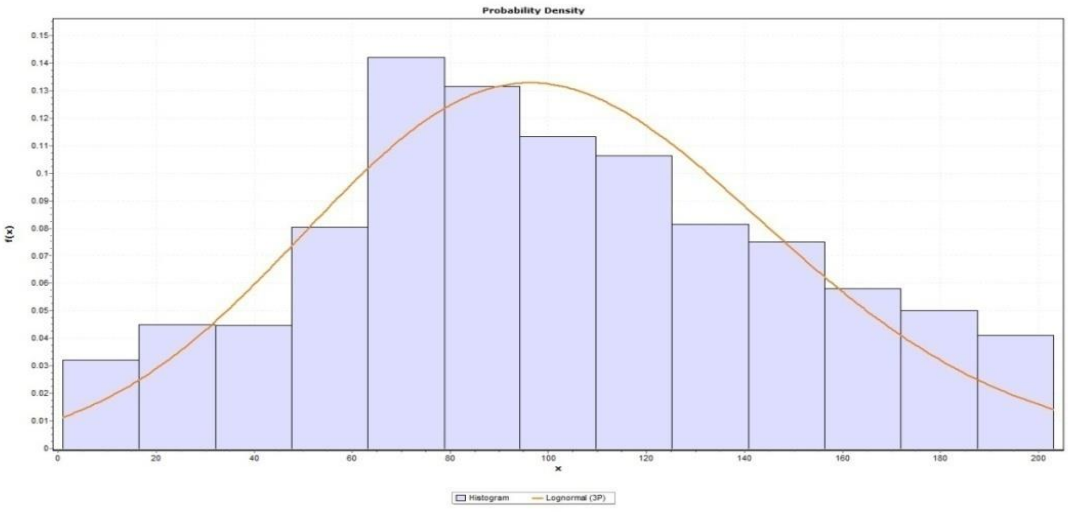


Figure 3.8 PDF for Other races (Lognormal distribution)

Table 3.7 Parameter estimates for the identified distributions

Race	PDF	Shape($\hat{k}$)	Location($\hat{\mu}$)	Scale($\hat{\sigma}$)
Whites	Generalized Extreme Value	-0.25296	81.455	49.931
African Americans	Generalized Extreme Value	-0.17371	67.907	49.663
Others	Log Normal	0.07439	-529.26	6.4442

3.5.2 Comparison of average survival and confidence interval estimation

The 95% confidence intervals for the mean duration and median survival for all the three race women are given below in Table 3.8.The median death time (median survival) for a White women with breast cancer is 179 months and for African American Women is 135 months. There is no median value reported for the survival of other race women because the product-limit estimator for these data never reached a failure probability greater than 42.40% or a survival probability lower than 57.60%. Now we proceed to identify the survival, hazard, cumulative hazard functions for the three races.

Let $\mu_w, \mu_{aa}, and\ \mu_{oth}$ represent mean survival times of whites, African Americans and other races respectively. Our interest is to test the hypothesis whether all the three races have same mean survival time or otherwise.

$$H_0: \mu_w\ = \mu_{aa}\ = \mu_{oth}\ vs. H_1:\ At\ least\ one\ of\ them\ is\ not\ equal.$$

By performing a one way analysis of variance at 5% level of significance, we obtained the p-value which is very low ($p < 0.0001$ for $F=96.413$); leading us to the conclusion that there is significant difference between the average mean survival times of the three races. Also, non-parametric testing using Kruskal-Wallis supports the current decision. So, we now proceed in

pair wise testing of mean survival times for all three races. The Table 3.9 below has the details of the results after performing t-test for pair wise testing. Clearly, we conclude that the average survival times is significantly different for all the three races in this study. Additionally, at 5% level of significance, we conclude that average survival times of White women is greater than African American women and less than other race women. African American women has less average survival compared to the other two races.

Table 3.8 Confidence intervals of mean duration and median survival

Race	Mean Survival: (95% CI)	Median survival: (95% CI)
Whites	100.05: (99.54, 100.56)	179: (175, 186)
African Americans	89.183: (87.61, 90.76)	135: (126, 145)
Others	101.5: (100.27, 102.73)	-

Table 3.9 Pair-wise hypothesis testing for average survival times of three races

H_{Null}	$H_{Alternative}$	P-value	Conclusion	95% CI for mean differences
$\mu_w = \mu_{aa}$	μ_w not equals μ_{aa}	0.000	Reject Null	(9.28, 12.45)
$\mu_{aa} = \mu_{oth}$	μ_{aa} not equals μ_{oth}	0.000	Reject Null	(-14.32, -10.33)
$\mu_w = \mu_{oth}$	μ_w not equals μ_{oth}	0.042	Reject Null	(-2.86, -0.05)

3.6 Cumulative Distributive Function

The cumulative distribution function is very useful in describing the continuous probability distribution of a random variable, such as time, in survival analysis. The cumulative distribution function (CDF) of a random variable T, denoted $F_T(t)$, is defined by $F_T(t) = P_T(T < t)$. This is interpreted as a function that will give the probability that the variable T will be less than or equal to any value t that we choose. Several properties ofa distribution function $F(t)$ can be listed as a consequence of the knowledge of probabilities. Because $F(t)$ has the

probability$0 < F(t) < 1$, then $F(t)$ is a non-decreasing function of t, and as t approaches ∞, $F(t)$ approaches 1. Figure 3.9, Figure 3.10, Figure 3.11 depict the respective CDF's for all the three races.

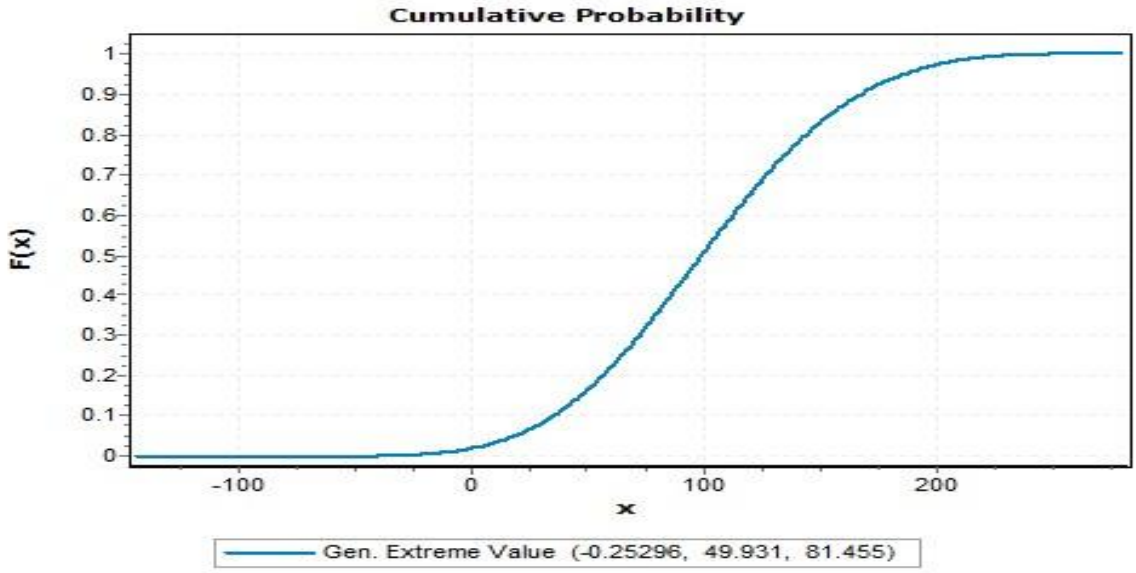


Figure 3.9 CDF for Whites

The fitted GEV CDF for the other White women is given below. The CDF graph is given in Figure 3.9.

$$F(x) = exp\left\{-\left[1 + (-0.253)\left(\frac{x - 81.455}{49.931}\right)\right]^{-1/(-0.253)}\right\}$$

In the Figure 3.9 above we can clearly notice that for White women with breast cancer the probability of surviving more than 100 months is little more than 50%. i.e. $P(\mathrm{X}_w > 100) = 0.5$. Thus, by 100 months, a White women identified with breast cancer has accumulated quite a bit of risk, which begins to accumulate more slowly after this point.

Similarly, from the Figure 3.10 below we can see that the probability of surviving 100 months or fewer is near 60%. i.e., $P(\mathrm{X}_{aa} > 100) = 0.4$. Thus, by 100 months, an African American women identified with breast cancer has accumulated quite a bit of risk, comparatively more than White women, which then begins to accumulate more slowly after this point.

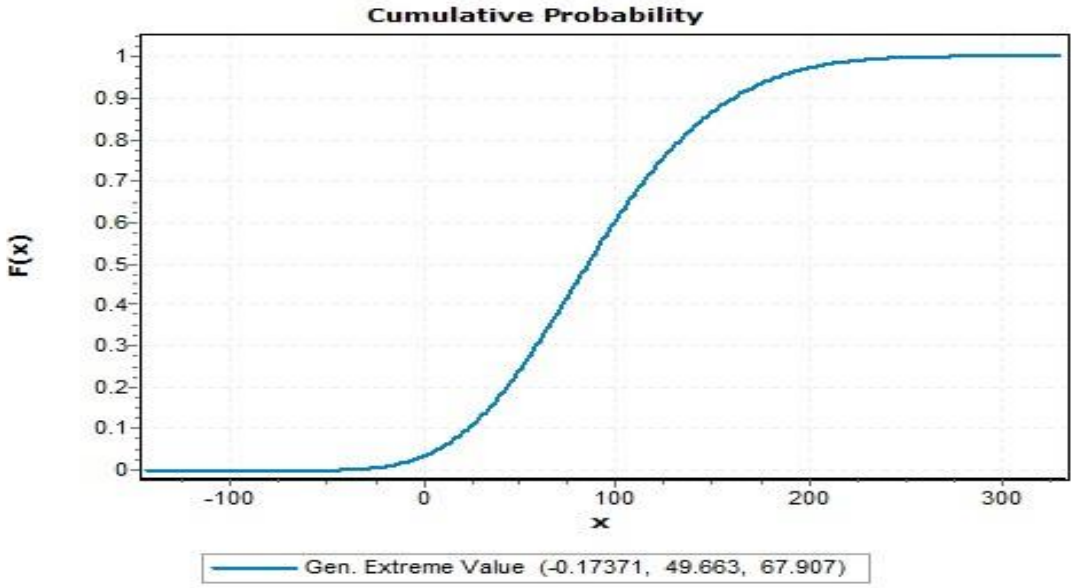


Figure 3.10 CDF for African Americans

The fitted GEV CDF for the other AA women is given below. The CDF graph is given in Figure 3.10.

$$F(x) = exp\left\{-\left[1 + (-0.174)\left(\frac{x - 67.907}{49.663}\right)\right]^{-1/(-0.174)}\right\}$$

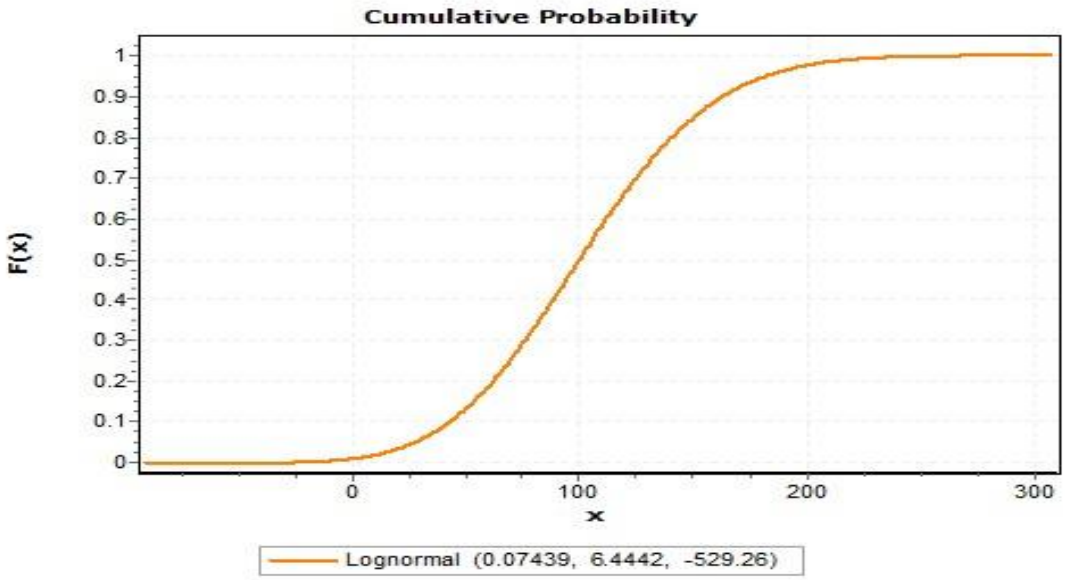


Figure 3.11 CDF for Others

The fitted Lognormal CDF for the other race women is given below. The graph of the same is given in Figure 3.11.

$$F(x) = \Phi\left\{\frac{\ln(x-(-529.26)) - 6.4442}{0.0744}\right\}$$

From the above Figure 3.11 the probability of surviving 100 months or fewer for other race women is near 50%. i.e., $P(\mathrm{X}_{oth} > 100) = 0.5$. Thus, by 100 months, equaling with White women survival, a patient from other races identified with breast cancer has accumulated quite a bit of risk by then.

3.7 Survival Function

Let T > 0 have a probability density function (PDF) $f(t)$and cumulative distribution function (CDF)$F(t)$. Survival experience is described by the cumulative survival function given by

Survivor function $S(t)$ = chance of surviving to age t

= percent still alive at age t

$$S(t) = P\{T > t\} = 1 - F(t)$$

Evidently, $S(t)$is the survival probability: the probability that the event will not happen until time *t*. The survival function gives the probability of surviving or being event-free beyond time t. Because $S(t)$ is a probability, it is positive and ranges from 0 to 1. It is defined as $S(0) = 1$ and as t approaches ∞, $S(t)$ approaches 0. The Kaplan-Meier estimator, or product limit estimator, is the estimator used by most software packages because of the simplistic step idea. The Kaplan-Meier estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. The survival curve describes the relationship between the probability of survival and time.

From the Figure 3.12, the probability of White women surviving beyond 150 months is a little less than 0.2, and we see that the probability of surviving 150 months or fewer is a little more than 0.8. From the Figure 3.13 and Figure 3.14, we notice that, the probability of African American women and other race women surviving beyond 150 months is a little less than 0.1 and 0.15 respectively. Clearly White women has more probability of survival than other two races. The fitted form of survival functions for all the three races are given below.

For white women: $S(x) = 1 - F(x) = 1 - exp\left\{-\left[1 + (-0.253)(\frac{x-81.455}{49.931})\right]^{-1/(-0.253)}\right\}$

For AA women: $S(x) = 1 - F(x) = 1 - exp\left\{-\left[1 + (-0.174)(\frac{x-67.907}{49.663})\right]^{-1/(-0.174)}\right\}$

For Other race women: $S(x) = 1 - F(x) = 1 - \Phi\left\{\frac{\ln(x-(-529.26))-6.4442}{0.0744}\right\}$

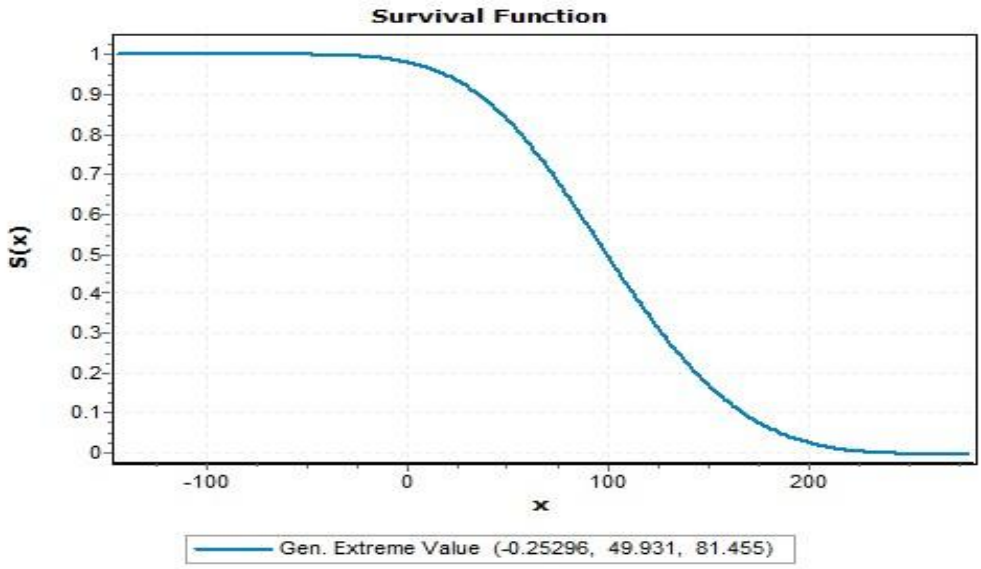


Figure 3.12 Survival DF for Whites

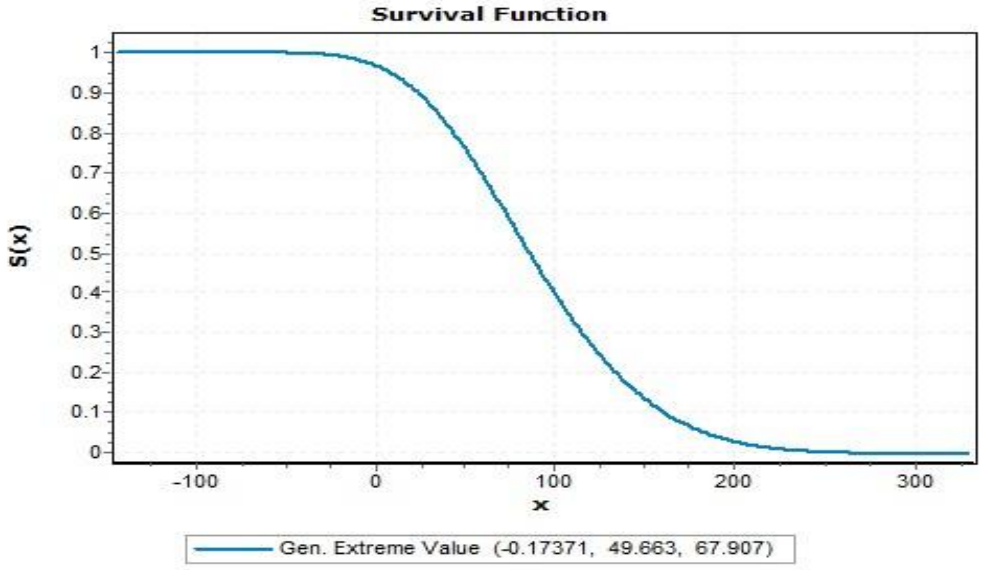


Figure 3.13 Survival DF for African Americans

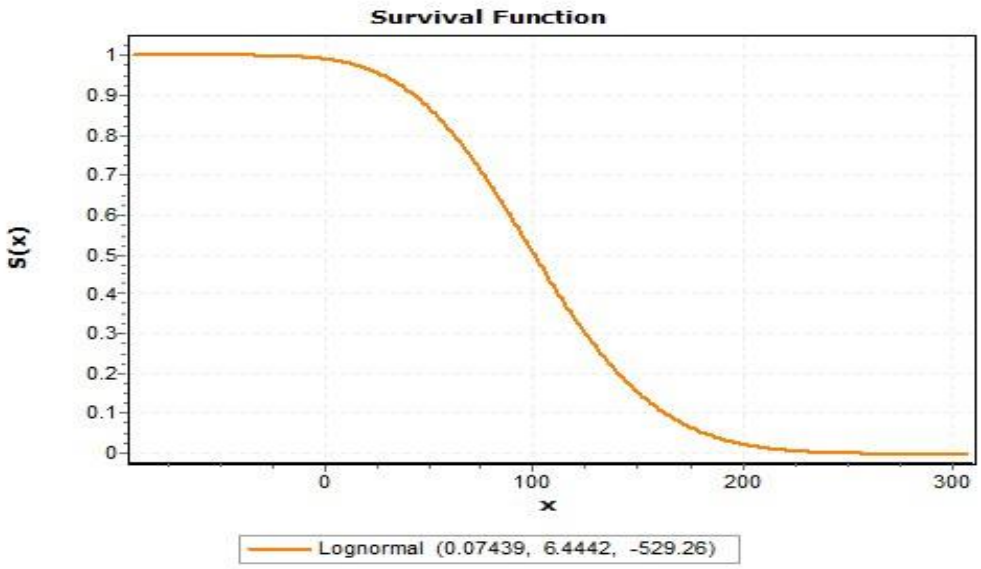


Figure 3.14 Survival DF for others

3.8 Hazard Function

The hazard at time *t*, $h(t)$ as the probability of an event at the interval $[t, t + \Delta t]$, when$\Delta t \to 0$. To find an expression for $h(t)$, we should realize that $h(t)$ must be a conditional probability: it is conditional on not having the event up to time *t* (or conditional on surviving to time *t*.). The hazard is the probability of dying (or experiencing the event in question) given that patients have survived up to a given point in time, or the risk for death at that moment.

The connection between hazard, survival, PDF and CDF is given below. The CDF is the best starting point. From CDF we get to PDF and then to hazard. Hazard function, $h(t)$ = age-specific death rate = percent dying at age t of those alive at age greater or equal to t,

$$h(t) = \frac{dF/_{dt}}{S(t)} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}$$

The hazard function has formulation as in the Cox model assumes the subject *i* at time *t* of the form, $h_i(t) = h_0(t)\ exp(X_i\beta^{PH})$, where X_i is the set of covariates for subject *i* (at time *t),* β^{PH} is the vector of fixed effects regression coefficients, $h_o\ (t)$ is the baseline hazard (at time *t*). The meaning of the formula stated above implies that if you survive to t, you will succumb to the event in the next instant. This function additionally assumes baseline h_0 to correspond to specific distribution with PH property. The above equation is the number of deaths per unit time in the interval divided by the average number of survivors at the midpoint of the interval. The hazard function is commonly known as the instantaneous failure rate. It is the measure of the risk of failure at a point in the time during the aging process.

The graph of the hazard rates of White women (Figure 3.15) shows that probability of failing (conditional on having survived to time t) remains below 0.05 for the first 100 months whereas from Figure 3.16 for African American Women probability of failing remains below 0.02 for the first 100 months and hazard rises steeply over 100 months. The hazard of other race women from Figure 3.17 displays the probability of failing is below 0.016 until first 100 months and then rising linearly thereafter. The fitted hazard function for all the three races are given below.

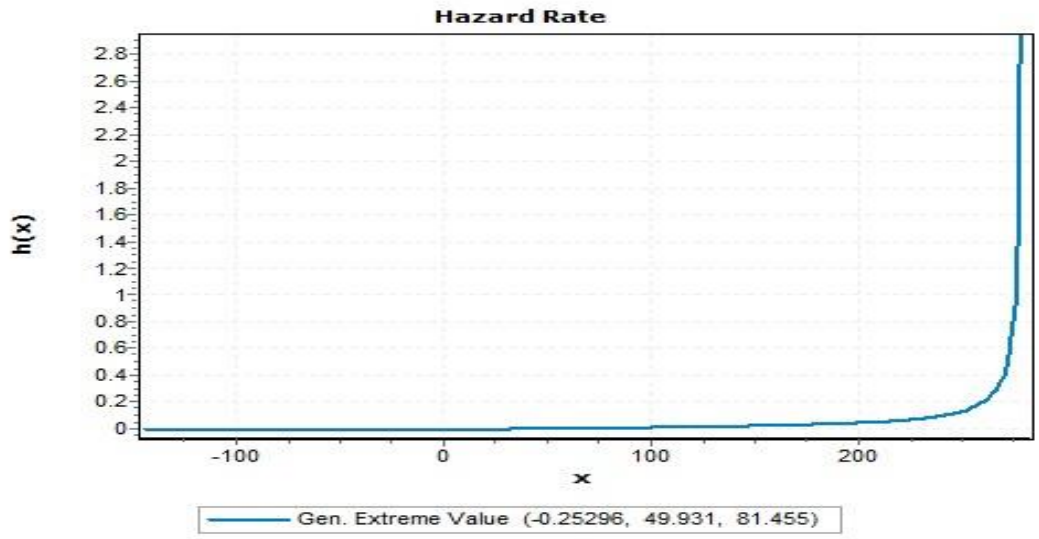


Figure 3.15 Hazard Function for Whites

$$h(x) = \frac{\frac{1}{49.931}\exp\{-(1+(-0.253)z)^{-1/(-0.253)}(1+(-0.253)z)^{-1-(-0.253)}\}}{1-exp\left\{-\left[1+(-0.253)(\frac{x-81.455}{49.931})\right]^{-1/(-0.253)}\right\}}$$

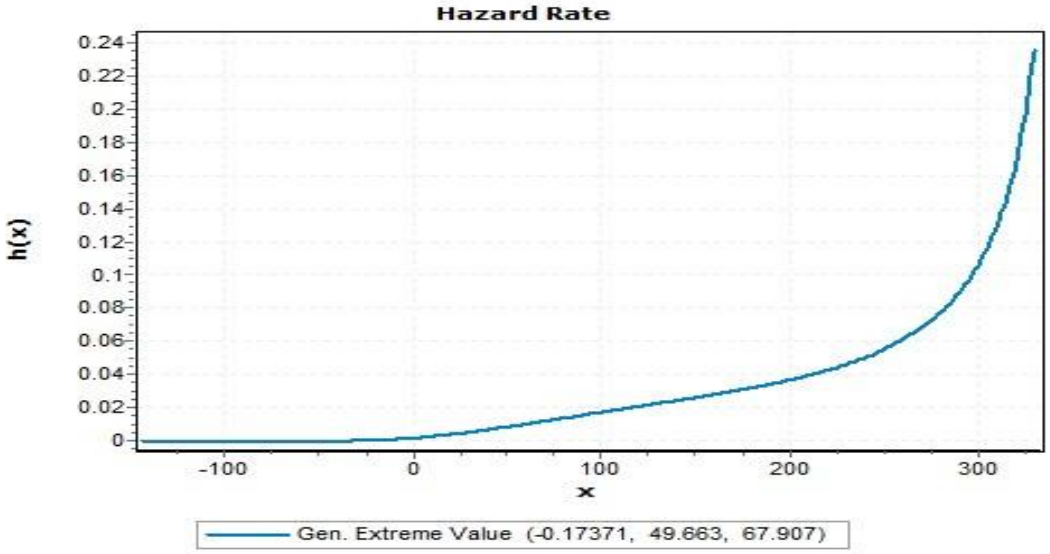


Figure 3.16 Hazard Function for African Americans

$$h(x) = \frac{\frac{1}{49.663}\exp\{-(1+(-0.174)z)^{-1/(-0.174)}(1+(-0.174)z)^{-1-(-0.174)}\}}{1-exp\left\{-\left[1+(-0.174)(\frac{x-67.907}{49.663})\right]^{-1/(-0.174)}\right\}}$$

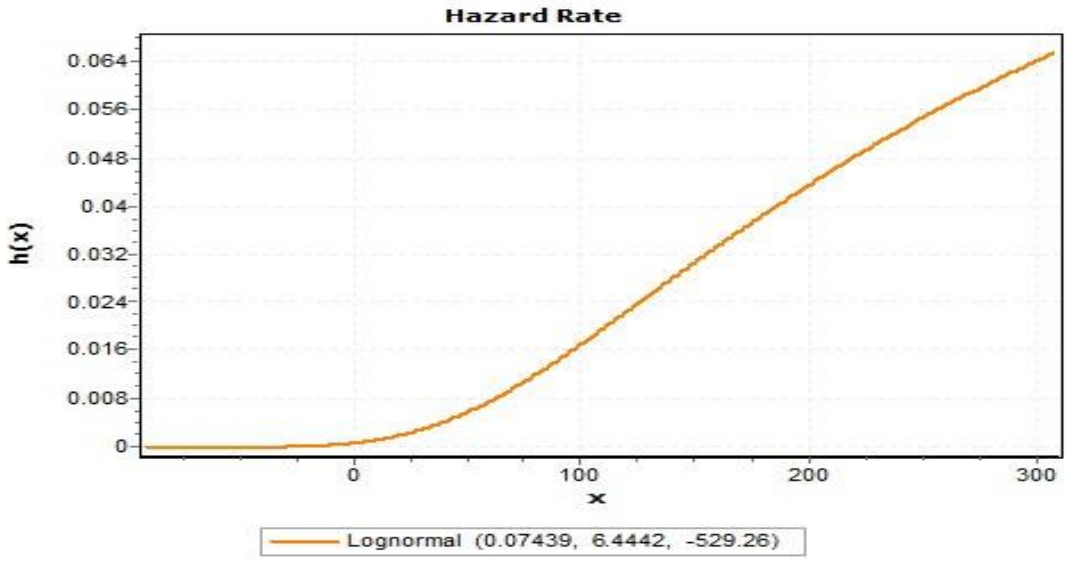


Figure 3.17 Hazard Function for Others

$$h(x) = \frac{\frac{1}{(0.0744)(x-(-529.26))\sqrt{2\pi}} exp\left\{-\frac{[\ln(x-(-529.26))-6.444]^2}{2(0.0744)^2}\right\}}{1-\Phi\left\{\frac{\ln(x-(-529.26))-6.4442}{0.0744}\right\}}$$

3.9 Cumulative Hazard Function

The cumulative hazard function $H(t)$is the integral of the hazard function $h(t)$. It can be interpreted as the probability of failure at time x given survival until time x. As the name implies, cumulative hazard function cumulates hazards over time.

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

Clearly is the area under the curve of the function $h(x)$, on the interval from 0 to t. A given cumulative hazard will remove a certain proportion of objects (or be associated with a probability of surviving beyond t). For example, a cumulative hazard of 0 (i.e.,$H(t) = 0$) has 100% associated survival (i.e.$S(t) = 1$). The above equation can also be expressed as $H(t) = -\ln(1 - F(t))$. The cumulative hazard function gives the number of expected number of failures

over time interval *t*. When the survival function is at its maximum at the beginning of analysis time, the cumulative hazard function is at its minimum. As time progresses, the survival function proceeds towards it minimum, while the cumulative hazard function proceeds to its maximum.

From Figure 3.18, Figure 3.19 and Figure 3.20 it is clear that the cumulative hazard function, H (t) increases more rapidly over time, supporting our previous results. H (t) for African Americans is comparatively more than the other two races.

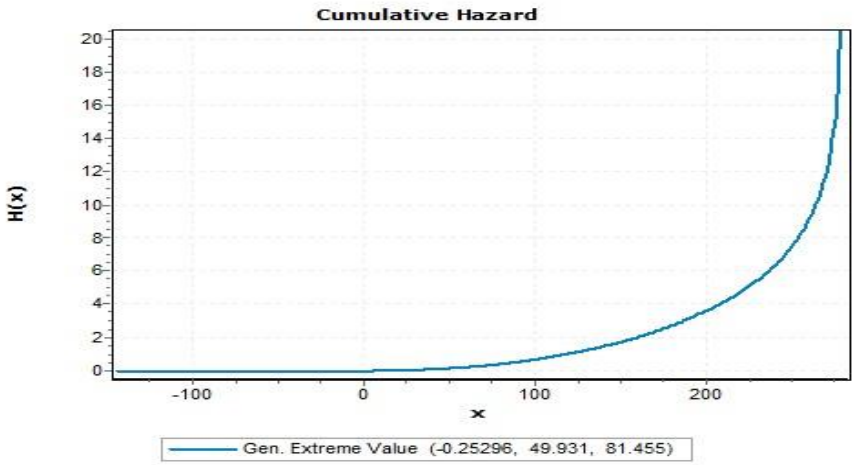


Figure 3.18 Cumulative Hazard Function for Whites

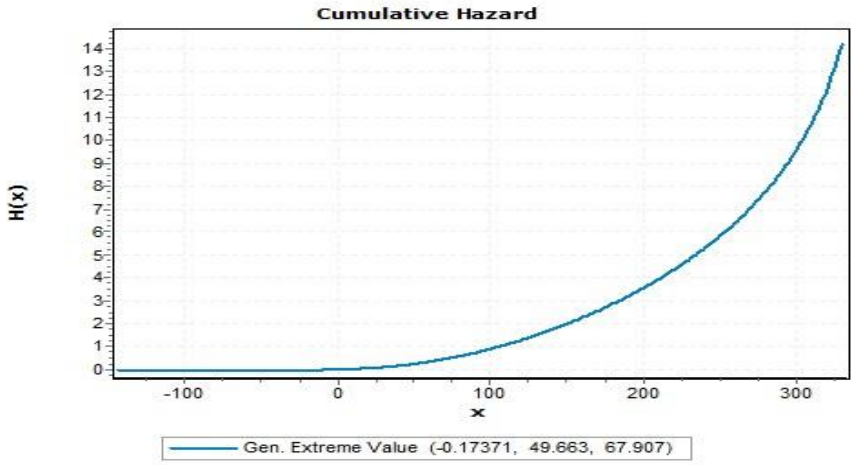


Figure 3.19 Cumulative Hazard Function for African Americans

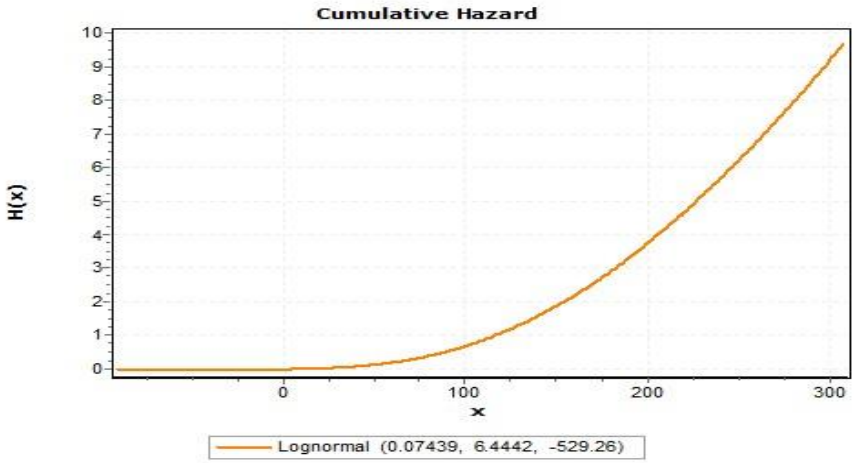


Figure 3.20 Cumulative Hazard Function for others

3.10 Conclusion

Lower p-values in log-rank test and product-limit survival curves indicated a statistically significant difference between the survival times of all the three races. Compared with White women and African American women, other race women has more probability of survival. This is also supported by survival curves and hazard functions in the later sections of this chapter. However the median survival for other race women and White women is almost same and African American women has comparatively very less median survival.

Survival resulting from breast cancer specifically were analyzed for the study population overall by race and treatment taken at diagnosis and summarized in Table 3.3 and Table 3.4. From Table 3.4, it is interesting to learn that the probability of survival and death is almost very close for the patients who underwent radiation alone. The probability density function that best characterizes the behavior of survival time are identified as GEV distribution for Whites and African American race women and Log Normal distribution for other race women. The parameter estimates of these distributions are given in Table 3.7.

CHAPTER FOUR

Modeling of Breast Cancer Survival Data

4.0 Introduction

Survival analysis today is widely implemented in the fields of medical and biological sciences, social sciences, econometrics, and engineering. The basic principle behind the survival analysis implies to a statistical approach designed to take into account the amount of time utilized for a study period, or the study of time between entry into observation and a subsequent event. The event of interest pertains to death and the analysis consisted of following the subject until death (36). Events or outcomes are defined by a transition from one discrete state to another at an instantaneous moment in time. Examples include time until onset of disease, time until stock market crash, time until equipment failure, and so on. Although the origin of survival analysis rests with the mortality tables from centuries ago, this type of analysis was not well developed until World War II (37). At the end of the war, the use of these newly developed statistical methods quickly spread through private industry as customers are demanding for safer and more reliable products.

In survival analysis, a data set can be categorized as exact or censored, and it may also be truncated. Another name for exact data is uncensored data which occurs only when the precise time until the event of interest is known. Censored data arises when a subject's time until the event of interest is known to occur only in a certain period of time. For example, if an individual drops out of a clinical trial before the event of interest has occurred, then that individual's time-

to-event is right censored at the time point at which the individual left the trial. The time until an event of interest is truncated if the event time occurs within a period of time that is outside the observed time period (38).

4.1 Questions of Interest

Q1: How long a woman with breast cancer will survive after undergoing certain treatments? (Radiation or surgery or both radiation and surgery or no treatment).

Q2: What is the effectiveness of treatments when implemented in different stages of breast cancer?

Q3: Given a vector of covariates or explanatory variables is there a parametric survival model that may affect survival time of breast cancer women?

Q4: How good is the popular Kaplan Meier survival analysis when compared with others (parametric and nonparametric functions)?

Q5: Does the Cox proportional hazards survival analysis provide any additional information with respect to survival function?

Q6: Is there any significant difference in proposed parametric survival model and Cox PH models?

4.2 Survival and Hazard functions

Survival time can be estimated as a variable which calculates the time between the starting point and ending point of event of interest or time of interest. In medical field (39) it is termed as the period elapsing between the completion or institution of any procedure and death. The survival time and event data is collected on practical grounds which is either censored or truncated.

Let us recall the definition of survival function as discussed in the previous chapter. This survival function is also termed as survivor or reliability function. It is defined as the probability associated with the mortality rate or failure of some system. This survival function (40) is obtained by plotting graph of associated probabilities against time. The survival function can be expressed with help of another distribution used commonly in statistical techniques, namely cumulative probability function CDF denoted as F(t). The survivor function is defined as the complement of the CDF which is formulated in the relationship below

$$S(t) \;=\; Pr(T > t) \;=\; 1 - \; F(t)$$

Similarly Hazard function is an alternative representation of the distribution of T or the instantaneous occurrence of the event and is defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t < T \leq t + dt | T > t)}{dt}$$

The above expression is termed as the instantaneous rate of occurrence for the conditional probability that the event will occur in the time interval between t and (t+dt) as it has not occurred before.

By the prior computation of the conditional probability in the numerator and application of limits gives the hazard function as

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

In other words the hazard function can be stated as the rate of the occurrence of the event at time t equals to the probability density at time t divided over the probability of the surviving to that duration without experiencing the event. The above formula can be expressed using the relation between density and survival function as follows

$$\lambda(t) = -\frac{d}{dt} \log S(t)$$

The above expression of hazard function is integrated using limits 0 to t and applying the boundary condition S(0)=1 (which implies event not occurred at time 0) to obtain relation between hazard and survival function as follows

$$S(t) = \exp(-\int_0^t \lambda(x)dx)$$

4.3 Statistical Approach of Survival Analysis

The survival analysis can be carried out using various statistical approaches (41) like

1. Descriptive statistics (includes mean or median of survival, average hazard rate etc.)
2. Univariate statistics (survival curves)
3. Multivariate statistics (Parametric, non-parametric and semi-parametric survival analysis)

The first two classifications of survival analysis have their respective advantages and disadvantages which are applicable in only few cases. The third classification is observed to be present generation scenario for survival function analysis. Survival models for the analysis of data have three main characteristics: (i) the dependent variable or response is the waiting time until the occurrence of a well-defined event, (ii) observations are censored, in the sense that for some units the event of interest has not occurred at the time the data are analyzed, and (iii) there are predictors or explanatory variables whose effect on the waiting time we wish to assess or control (128).

The basic definition of three types of analysis carried under multivariate statistics are given below.

1. Parametric Analysis: This analysis assumes distributions for outcome, and base statistical analysis on assumed distributions (check the validity of assumptions).

2. Non-parametric Analysis: This analysis avoids distribution or quantitative assumptions and relies completely on design properties.
3. Semi-parametric Analysis: This analysis is an intermediate between above two types of analysis, but will make some assumptions to avoid fully specified statistical model (42).

4.4 Non-parametric Analysis (NP)

Estimating the distribution of the dependent variable without making assumptions about its shape is an important first step in analyzing a dataset. Given the importance of the distribution of the dependent variable it is valuable to "let the data speak for itself" first (43). Estimating the probabilities without making any assumptions on its shape is called non-parametric analysis. The function used to represent the distribution is the Survivor function. Nonparametric methods do not require the knowledge of the underlying distribution of the failure time t. Hence it provides an edible way to deal with the data in many practical situations. The seminar paper by Kaplan and Meier (44) is the benchmark in survival analysis especially from nonparametric point of view. It compelled the application of descriptive statistics and improved the development of all existing NP approaches with censored data. The survivor function is calculated by dividing the number of survivors by the total number of subjects for every time.

4.4.1 Kaplan-Meier Estimator

The Kaplan-Meier estimator originally was derived as an NP maximum likelihood estimator of $F(t)$. Because of the latter method of derivation, it is also called as the product-limit (PL) estimator. If the data was not censored then the empirical survival function is given by

$$S(t) = \frac{1}{n}\sum_{i=1}^{n} I\{t_i > t\},$$

Where I is termed as the indicator function which takes a value of one if the condition $t_i > t$ is true or zero otherwise (45). The estimator is simply the proportion alive at t. For the censored data, assume the ordered times of death as$t_1 < t_2 < t_3 \ldots \ldots < t_m$ and d_k be the death occurred at t_k. Let n_k be the number of persons alive just before t_k. This is the number exposed to risk at that time. The Kaplan-Meier (KM) or product limit estimate of the survivor function is

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right)$$

The justification of the estimate is explained as follows. In order to survive until the time t one must first survive until the time t_1. And the conditional probability of surviving from t_2 to t_1 given already survived t_1 is to be satisfied. The Kaplan-Meier (KM) is a step function with jumps at the observed times. If no censoring is present, the KM coincides with the empirical survival function (46).

As mentioned earlier, KM estimator can be interpreted as the non-parametric likelihood estimator (NPML) for the death or censored data at time t. The assumptions formulated for this method requires that the likelihood of the subject is $S(t)$ at t is to be maximized as large as possible. Since the survival is a non-decreasing function, it does not change at the censoring times. Also if a person dies at t which is distinct from times of the death we introduced before. Let it be time t_i. We need to make the survival function before t_i as large as possible. Based on the above criteria the likelihood takes the form

$$L_i = \prod_{i=1}^{m} [s(t_{(i-1)}) - S(t_{(i)})]^{d_i} S(t_{(i)})^{c_i}$$

This is the product over m distinct times of death. By taking $t(0) = 0$ with $S(t(0)) = 1$. Estimation of survival function at the death times $t(1), t(2), \ldots, t(m)$ for m parameters is obtained.

$$\pi_i = \frac{S(t_i)}{S(t_{i-1})}$$

Writing the above expression for the conditional probability of surviving from $S(t_{i-1})$ to $S(t_i)$. Then we can write $S(t_i) = \pi_1\pi_2 \ldots\ldots\ldots \pi_i$, and the likelihood changes to the following expression

$$L_i = \prod_{i=1}^{m}(1-\pi_i)^{d_i}\pi_i^{c_i}(\pi_i\pi_2 \ldots\ldots\ldots \pi_{i-1})^{d_i+c_i}$$

In all these cases, individuals who die at time $t_i$or the time between t_i and t_{i+1}also contribute to the term π_j to each of the previous term of death from $t_{(1)}$to $t_{(i-1)}$. Let us assume $n_i = \sum_{j\geq i}(d_j + c_j)$ to be number exposed to risk at $t_i$and now the L likelihood can be written as

$$L_i = \prod_{i=1}^{m}(1-\pi_i)^{d_i}\,{\pi_i}^{n_i-d_i}$$

The maximum likelihood estimator of π_i is then

$$\hat{S}(t) = \hat{\pi}_i = \frac{n_i - d_i}{n_i} = 1 - \frac{d_i}{n_i}.$$

In order to estimate $var\left(\hat{S}(t)\right)$, we use the delta method which says, if X $\sim N(\mu, \sigma^2)$ then $f(X)$ is approximately normally distributed with mean $f(\mu)$ and variance $[f'(\mu)]^2\sigma^2$. Also instead of estimating the $var\left(\hat{S}(t)\right)$, we can use the delta method to approximate the $var(\log\left(\hat{S}(t)\right)$ with $\log\left(\hat{S}(t)\right)$=$\sum_{j:t_j<t}\log(1-\hat{\lambda}_j)$. Using independence of the $\hat{\lambda}_j's$ we get the Greenwood's Formula given by

$$var\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 var\left[\log\left(\hat{S}(t)\right)\right],$$

Implying,

$$var\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{(r_j - d_j)r_j}$$

And the sample standard error for computing confidence interval is given by

$$SE\left(\hat{S}(t)\right) = \hat{S}(t)\sqrt{\sum_{j:t_j<t} \frac{d_j}{(r_j - d_j)r_j}}$$

4.4.2 The Nelson-Aalen Estimator

For estimating a cumulative hazard $H(t)$, one simple approach is to find an estimator of $S(t)$ and take minus the log. An alternativeapproach is to estimate the cumulative hazard directly using the Nelson-Aalen estimator. The Nelson Aalen estimator is a non-parametric estimator of the cumulative hazard rate function from censored survival data (47). Consider a sample of *n* individuals from a right censored survival population. Our observation of the survival times for these individuals will typically be subject to right censoring meaning that for some individuals we only know that their true survival times exceed certain censoring times. The censoring is assumed to be independent in the sense that the additional knowledge of censorings before any time *t* does not alter the risk of failure at *t*. The Nelson-Aalen estimator is a step function with the location of the steps placed at each observed death time and the vertical size of the steps is the inverse of number at risk, Where number at risk is the number of patients just before the death that are still observed to be alive. With larger samples the Nelson-Aalen estimator will get closer to the true cumulative hazard. The Nelson-Aalen estimator is given by

$$\widehat{H(t)} = \sum_{t_j<t} \frac{d_j}{r_j}$$

where d_j is the subjects who die at time$t_j$and $r_j$is the number of subjects at risk just prior to time t_j. The variance of the estimator is given by

$$Var(\widehat{H(t)}) = \sum_{t_j<t} \frac{d_j(r_j - d_j)}{{r_j}^2(r_j - 1)} \approx \sum_{t_j<t} \frac{d_j}{{r_j}^2}$$

The advantage of non-parametric analysis is that the results do not rest on the assumptions. The disadvantage is that we can only compare limited number of groups which implies it is very difficult to see the impact of multiple explanatory variables on the subjects (48). Another disadvantage of non-parametric techniques is that it can only deal with the quantitative explanatory variables like GDP, rich and poor countries etc.

4.4.3 Kaplan Meier Estimation for breast cancer survival

4.4.3.1 Effect of treatments on survival of breast cancer

Considering the breast cancer survival data, in this chapter we are interested in knowing how long women with breast cancer will survive after undergoing certain treatments. Treatments include radiation or surgery or both radiation and surgery or no treatment. Also we would like to know the effectiveness of treatments when implemented in different stages of breast cancer. Firstly we considered the effectiveness of treatments on survival for the overall data. From the Table 4.1 women who are treated with radiation have a median survival of 154 months with 95% CI (149, 157) months. Interestingly, women those who are treated with both treatments has the same median survival as of women who received surgery. There is no median value reported for the survival of women who did not receive any treatment because the KM estimator for these data never reached a failure probability greater than 41.50% or a survival probability lower than

58.50%. Figure 4.1 is the product-limit survival graph for all the four treatment types. Treatments 3 & 4 in the graph follow almost the same path. The probability of survival for a women identified with breast cancer to survive more than 50 months is approximately 82% for women who did not receive any treatment, 78% for surgery, 30% for combination of radiation and surgery and 30% for those who are treated with surgery.

Table 4.1 Treatment wise KM estimates for median survival

Treatment	No treatment	Radiation	Radiation & Surgery	Surgery
Median Survival	-	154	25	25
95% CI	-	[149, 157)	[21, 30)	[21, 29)

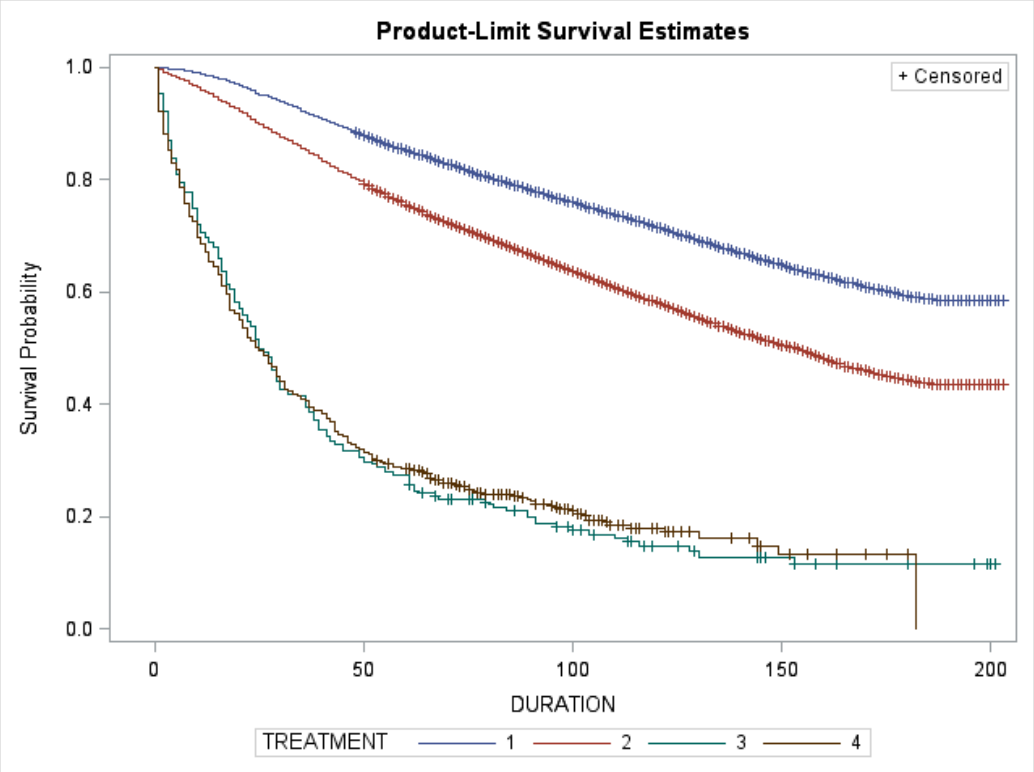


Figure 4.1 Product-Limit estimates for treatments

Table 4.2 Stage vs. Treatment Product-Limit Estimates for median survival

	No treatment	Radiation	Radiation & Surgery	Surgery
Stage I	—	178(172, -)	104(53, -)	62(42, 100)
Stage II	—	145(140, 149)	128(45, -)	43(32, 66)
Stage III	93(81, 103)	52(47, 57)	32(24, 110)	27(17, 31)
Stage IV	34(28, 39)	23(21, 27)	17.5(12, 22)	14.5(11, 18)

4.4.3.2 Stage wise effect of treatments of breast cancer

Further we continued to check the survival probability of breast cancer women treated in every stage with different treatments. The median survival in months and the respective 95% confidence interval based on their stage of cancer is tabulated in Table 4.2. The median survival for those who are in stage I and stage II who did not receive any treatment for breast cancer is not reported because the KM estimator for these data never reached a failure probability greater than 35.28% or a survival probability lower than 64.72% for the former case and data never reached a failure probability greater than 45.29% or a survival probability lower than 54.71% for the latter case. The Figure 4.2 clearly depicts that the probability of survival for those who are treated with surgery in all the four stages falls down rapidly.

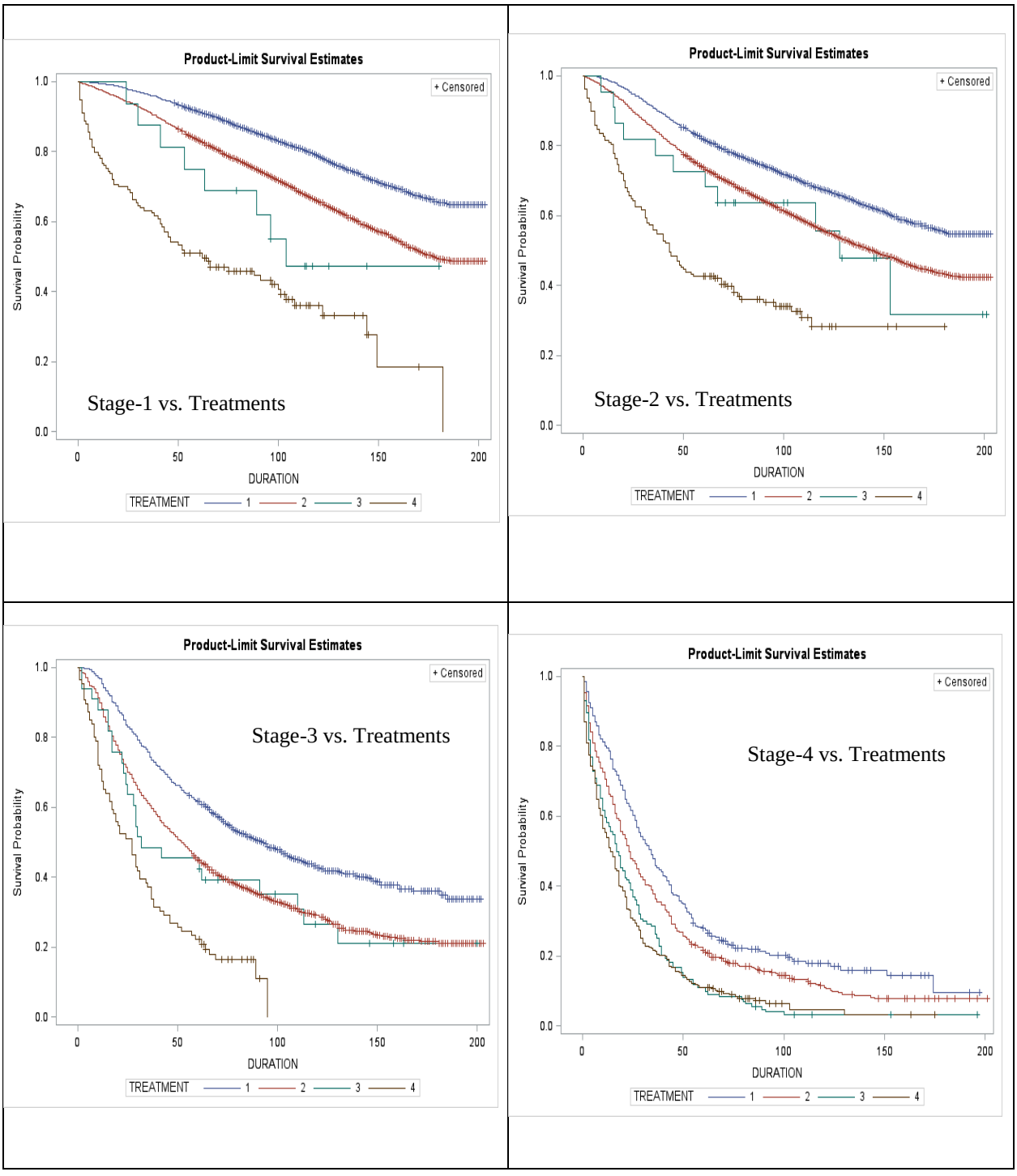


Figure 4.2 KM Estimates for Stages Vs. Treatments

4.5 Parametric Analysis

This type of analysis assumes a functional form of the probability distribution and the way in which explanatory variables influence the survival time. The first assumption is also called as time dependence because of its functional form of probability distribution. With growing computing power and existing statistical programming languages, it is relatively simple to work with exact likelihood for censored or truncated data with a variety of parametric models. Parametric survival model provides the possibility of more efficiency (43). It is proved to be interesting to assume a specific distribution for underlying hazard function (to obtain a full hazard or survival function). There may be a provision for non-proportional hazard functions also. The direct regression approach for the survival time estimation is given by

$$E(t_i) = \beta_o + X_i\beta \;\; or \;\; t_i = \beta_o + X_i\beta + \varepsilon_i$$

where ε_i refers to the survival error distribution.

This direct regression computation has some problems for estimating survival time like the distribution of t_i is right skewed (non-normal), the estimator of time may not be the parameter of interest (not equal to hazard) and censoring must be accounted.

The above concerns are addressed using two possible approaches of parametric analysis.

1. Accelerated failure time models (AFT models)

2. Proportional hazard model (PH models) (49)

These models are provided as the common scales for the distributions in parametric survival models. Both PH and AFT models were analyzed on basis of t-scale over the distributions with interval (0, ∞), whereas the AFT models were also interpreted on the basis of $ln(t) -$ scale over the distributions termed as pure AFT models. Distributions that are commonly used in parametric analysis using AFT are addressed below.

4.5.1 Parametric Model selection: Goodness of fit Tests

There are few common statistical methods for comparisons of survival models.

a) Log-likelihood test for the censored data,

b) AIC,

c) Cox-Snell Residual plots and

d) Likelihood-Ratio Statistic.

The AIC is an operational way of trading off the complexity of an estimated model against how well the model fits the data. The AIC is calculated by

$$AIC = -2\,log\,(likelihood) + 2\,(p + k),$$

where p is the number of parameters, $k = 1$ for the exponential model, $k = 2$ for the Weibull, log logistic, and log normal models and $k = 3$ for generalized gamma.

A likelihood ratio test (LRT) is also used to compare the fit of two models. The LRT test statistic is twice the difference in the log-likelihoods of the models considered for comparison. We generally select the model that gives the largest log-likelihood.

Other methods include graphical methods (for all distributions mentioned), Cox-Snell Residual plots among others (52). Parametric models are fit to the event times and semi-parametric models are fit to the ordered event times respectively. In both the cases we use the AIC to select between parametric models, or to select between semi-parametric models, but not to select from a mixture of the two. The AIC or likelihood tests allow us to assess relative model goodness of fit, but not absolute model goodness of fit. Just because the second model fits better than the first model, it does not mean the second model adequately describes the data. Thus, we

would like a method, at least a graphical one that lets us assess the absolute goodness of fit of a parametric model. The Table 4.3 below provides information regarding graphical check for goodness of fit for the identified parametric model for survival data (53, 54).

Table 4.3 Graphical check for goodness of fit for parametric survival models

Graph	Behavior	Resulting Distribution
$-logS(t)$ versus t	Straight line through origin.	Exponential
$\log[-logS(t)]$ versus $\log t$	Straight line	Weibull
$\Phi^{-1}(1-S(t))$ versus $\log t$	Straight line, where $\Phi(\,)$ is the CDF.	Log-normal
$\log\left[\frac{1-S(t)}{S(t)}\right]$ versus $\log t$	Straight line	Log-logistic

4.5.2 Parametric modeling of breast cancer data

Our data consists of 47167 breast cancer patients identified with malignant breast tumors. Patients are either White women, African American women or other race women, stratified into four stages of cancer and are treated with either radiation or surgery or combination of both or no treatment. Other covariates include grade of tumors, number of primary tumors, age, and marital status. For the rest of this chapter the variables and their representations are given in the Table 4.4 below.

It is of substantial interest in performing the parametric modeling is to see the difference in survival (in months) between those patients undergone with different treatments, after adjusting for patient's cancer stage, age, marital status, race, grade of tumor, and the number of primary tumors. We used SAS software to fit different parametric models. After performing univariate analysis marital status of woman is not statistically significant and hence is dropped from modeling.

When comparing parametric models, the Akaike Information Criterion (AIC) and log-likelihood values (54) can be used to select the best parametric model. The best fit model is the one with smaller AIC and largest Log-likelihood. Once the model is identified we will perform a residual analysis check that lets us assess the absolute goodness of fit of the identified parametric model.

Table 4.4 Variables used in survival modeling

Age	X_1
Grade	X_{2i} i=1:Well differentiated 2:Moderately differentiated 3:Poorly differentiated 4:undifferentiated 9:Cell type not determined (reference)
Numprims	X_3
Treatments	X_{4i} i=1:No Treatment 2:Radiation 3: Radiation & Surgery 4: Surgery (reference)
Stage	X_{5i} ; i=1,2,3,4(reference)
Race	X_{6i} i=1: Whites 2: African Americans 3: Other races(reference)
Tumor Size	X_7

4.5.3 Parametric survival model using AFT class

Let $T_i$denote a continuous non-negative random variable representing survival time of the i^{th} unit, the logarithm can be used as conventional modeling which is formulated below

$$\ln(T_i) = X_i\beta + \sigma\epsilon_i \;\; or \;\; T_i = \exp(X_i\beta)\exp(\sigma\varepsilon_i) \; = T_{0i}\exp(X_i\beta)$$

where ε_i is termed as a suitable error in the ln(t)-scale which is specific for a distribution, and $T_{0i} = \exp(\sigma\varepsilon_i)$ is the error corresponding to the original (t) scale. The term T_{0i}indicate the baseline function at $i = 0$. This implies that the explanatory variables act in multiples and direct product on the survival time and their effect is to increase or decrease the time of death with respect to the baseline function. The baseline function is specified up to an unknown parameter. The term $\exp(-X_i\beta)$ is termed as the acceleration parameter. This parameter is different from the value in PH model. From the industrial applications point of view, the name 'accelerated life' implies to the testing of the units to substantial worse conditions rather than they actually encounter in real life. Different kinds of parametric models are obtained assuming different types of distributions for error termϵ_i.Accelerated life models are considered as standard regression models appliedto the natural logarithm of survival time, and except for the fact that observations are censored, pose no new estimation problems. This model estimates goodness of fit for different distributions using Likelihood ratio (LRT) or Akaike Information criterion (AIC). Once the distribution of the error term is chosen, estimation is carried out by maximizing the log-likelihood for censored data (50) which is also termed as a Tobit model in economic literature.

4.5.4 Exponential distribution

In regression models it is common practice that the dependent variable depends on the explanatory variables only through a linear function. Because of its historical significance, mathematical simplicity and important properties, the exponential distribution is one of the most popular parametric models. This is the simplest possible distribution with one parameter which is derived treating the hazard function as a constant and of monotonic value over baseline hazard function denoted as $h(t) = \lambda$.

$$h(t) = \exp(\beta_0)\exp(X_i\beta)$$

So for the exponential distribution the instantaneous failure rate is independent of t so that the conditional chance of failure does not depend on how long the individual has been on trial. This is referred to as the memory less property of the exponential distribution.

4.5.4.1 Fitting Exponential Model

The survival and fitted survival functions for exponential parametric model are given by equations below. Table 4.5 has the analysis of the maximum likelihood estimation of parameters of the Exponential model for breast cancer patients.

$S(t;\boldsymbol{X}) = \exp(-t[exp(-b_0 - b_1X_1 - b_2X_2 \ldots - b_kX_k)])$ and

$S(t;\boldsymbol{X}) = exp\{-t\{\exp(-5.36 + 0.03X_1 - 0.32X_{21} - 0.14X_{22} + 0.17X_{23} + 0.21X_{24} + 0.09X_3 - 0.95X_{41} - 0.63X_{42} - 0.26X_{43} - 1.89X_{51} - 1.46X_{52} - 0.72X_{53} + 0.15X_{61} + 0.40X_{62} + 0.0004X_7)\}\}$.

Table 4.5 Analysis of MLEs for Exponential Model

Analysis of Maximum Likelihood Parameter Estimates							
Parameter		**DF**	**Estimate**	**S. E.**	**95% Confidence Limits**		**Pr > ChiSq**
Intercept		1	5.3589	0.0831	5.1960	5.5219	<.0001
AGE		1	-0.0321	0.0006	-0.0334	-0.0309	<.0001
GRADE	**1**	1	0.3171	0.0350	0.2485	0.3857	<.0001
GRADE	**2**	1	0.1368	0.0293	0.0794	0.1943	<.0001
GRADE	**3**	1	-0.1723	0.0291	-0.2293	-0.1153	<.0001
GRADE	**4**	1	-0.2108	0.0560	-0.3206	-0.1011	0.0002
GRADE	**9**	0	0.0000	.	.	.	.
NUMPRIMS		1	-0.0851	0.0135	-0.1116	-0.0586	<.0001
TREATMENT	**1**	1	0.9496	0.0509	0.8499	1.0494	<.0001

Table 4.6 (Continued) Analysis of MLEs for Exponential Model

Analysis of Maximum Likelihood Parameter Estimates							
Parameter		**DF**	**Estimate**	**S. E.**	**95% Confidence Limits**		**Pr > ChiSq**
TREATMENT	**2**	1	0.6257	0.0499	0.5278	0.7236	<.0001
TREATMENT	**3**	1	0.2614	0.0869	0.0910	0.4318	0.0026
TREATMENT	**4**	0	0.0000	.	.	.	.
STAGE	**1**	1	1.8855	0.0381	1.8108	1.9602	<.0001
STAGE	**2**	1	1.4624	0.0373	1.3894	1.5354	<.0001
STAGE	**3**	1	0.7280	0.0418	0.6462	0.8099	<.0001
STAGE	**4**	0	0.0000	.	.	.	.
RACE	**1**	1	-0.1512	0.0265	-0.2033	-0.0992	<.0001
RACE	**2**	1	-0.3980	0.0337	-0.4642	-0.3319	<.0001
RACE	**3**	0	0.0000	.	.	.	.
TUMOR_SIZE		1	-0.0004	0.0001	-0.0005	-0.0002	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000	
Weibull Shape		0	1.0000	0.0000	1.0000	1.0000	

4.5.4.2 Exponential Residual Plot

To evaluate the goodness of fit for exponential model we performed a residual analysis for observed and fitted data. The result shows that the mean residual is 0.3785, with a standard deviation of 0.3523 and a range of 12.157. A residual graph of survival functions for exponential parametric model is shown in Figure 4.3. Clearly the fitted data does not fall close to the straight line which explains that exponential is not the best fit for this data.

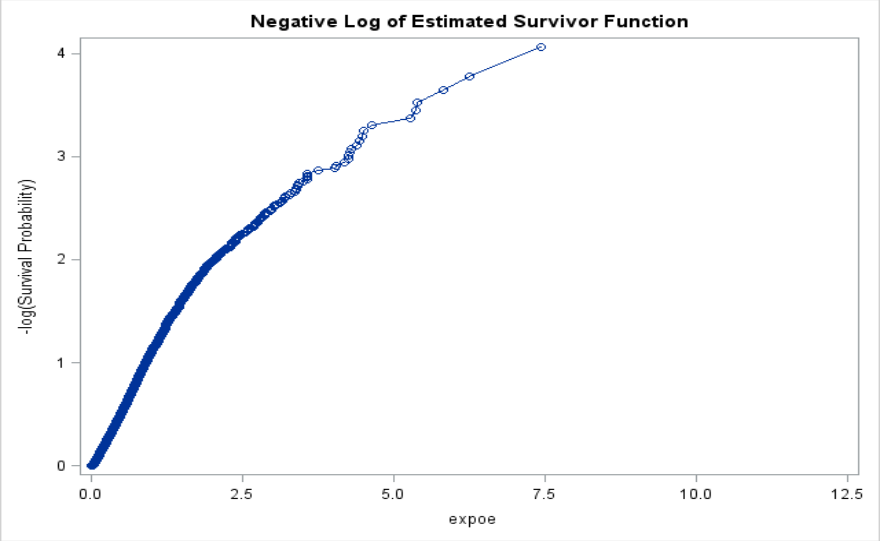


Figure 4.3 Residual plot for exponential distribution

4.5.5 Weibull distribution

Although the exponential model is good, there is improper assumption that the hazard function is constant over the time. If the hazard model is increasing or decreasing over the time, the exponential model will miss this fact under such assumption. The general Weibull model as a hazard function can be formulated as

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

The parameter p is called as the shape parameter which is one in case of exponential distribution. For the values of p other than one the hazard function increases or decreases monotonically. In case of AFT, Weibull model is represented as,

$$T_i = \exp(X_i\beta) \times \sigma\epsilon_i$$

Which implies the shape function is determined by the variance of the residuals. Intuitively, data with low variance duration dependence will tend to exhibit positive duration dependence, due to their relative lack of heterogeneity. Furthermore, Weibull in case of hazard ratio for two observations with different values *i* and *j* can be interpreted as follows

$$HR_{\frac{i}{j}} = \frac{\exp(X_i\beta)}{\exp(X_j\beta)}$$

This indicates that the hazard ratio for different cases can only differ by dichotomous variable which is $\exp(\beta)$.

4.5.5.1 Fitting Weibull Model

The survival and fitted survival functions for Weibull parametric model are given by equations below. Table 4.6 has the analysis of the maximum likelihood estimation of parameters of the Weibull model for breast cancer patients.

$S(t;\boldsymbol{X}) = \exp(-t^k exp(-b_0 - b_1X_1 - b_2X_2 \ldots - b_kX_k))$ and

$$\begin{aligned} S(t;\boldsymbol{X}) = exp\{-t^{0.85}\{\exp(5.22 &+ 0.03X_1 - 0.25X_{21} - 0.10X_{22} + 0.20X_{23} + 0.21X_{24} \\ &+ 0.07X_3 - 0.88X_{41} - 0.59X_{42} - 0.24X_{43} - 1.7X_{51} - 1.32X_{52} - 0.67X_{53} \\ &+ 0.13X_{61} + 0.35X_{62} + 0.0003X_7)\}\} \end{aligned}$$

4.5.5.2 Weibull Residual Plot

To evaluate the goodness of fit for the Weibull model we performed a residual analysis for observed and fitted data. The result shows that the mean residual is 0.3785, with a standard deviation of 0.3887 and a range of 14.633. A residual graph of survival functions for Weibull parametric model is shown in Figure 4.4. Clearly the fitted data does not fall close to the straight line which explains that exponential is not the best fit for this data.

Table 4.7 Analysis of MLEs for Weibull Distribution

Analysis of Maximum Likelihood Parameter Estimates							
Parameter		**DF**	**Estimate**	**S. E.**	**95% Confidence Limits**		**Pr > ChiSq**
Intercept		1	5.2216	0.0708	5.0827	5.3604	<.0001
AGE		1	-0.0287	0.0006	-0.0297	-0.0276	<.0001
GRADE	**1**	1	0.2513	0.0298	0.1929	0.3097	<.0001
GRADE	**2**	1	0.0974	0.0249	0.0486	0.1462	<.0001
GRADE	**3**	1	-0.1761	0.0247	-0.2245	-0.1277	<.0001
GRADE	**4**	1	-0.2086	0.0475	-0.3017	-0.1155	<.0001
GRADE	**9**	0	0.0000	.	.	.	.
NUMPRIMS		1	-0.0694	0.0115	-0.0919	-0.0469	<.0001
TREATMENT	**1**	1	0.8678	0.0433	0.7829	0.9527	<.0001
TREATMENT	**2**	1	0.5897	0.0424	0.5065	0.6728	<.0001
TREATMENT	**3**	1	0.2360	0.0738	0.0914	0.3805	0.0014
TREATMENT	**4**	0	0.0000	.	.	.	.
STAGE	**1**	1	1.6954	0.0331	1.6305	1.7603	<.0001
STAGE	**2**	1	1.3236	0.0320	1.2608	1.3863	<.0001
STAGE	**3**	1	0.6660	0.0355	0.5964	0.7356	<.0001
STAGE	**4**	0	0.0000	.	.	.	.
RACE	**1**	1	-0.1255	0.0225	-0.1696	-0.0813	<.0001
RACE	**2**	1	-0.3465	0.0287	-0.4028	-0.2903	<.0001
RACE	**3**	0	0.0000	.	.	.	.
TUMOR_SIZE		1	-0.0003	0.0001	-0.0005	-0.0002	<.0001
Scale		1	0.8484	0.0055	0.8376	0.8593	
Weibull Shape		1	1.1787	0.0077	1.1638	1.1939	

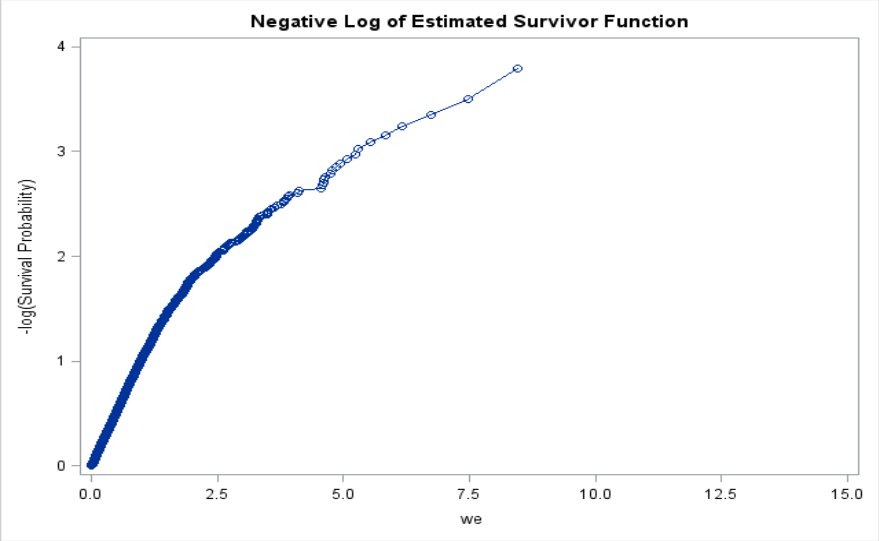


Figure 4.4 Residual plot for Weibull distribution

4.5.6 Log-normal and Log-Logistic distributions

There are also certain other models which received importance in social sciences. They can be noted probably as common models which are beyond the exponential and Weibull models. Both of the models are considered strictly AFT models. Recalling the general equation of AFT model, $\ln(T_i) = X_i\beta + \sigma\epsilon_i$.

If the error ϵ_i in the above equation is assumed to follow a logistic distribution (55) then the resulting model is termed as the log-logistic survival model. If the model follows a standard normal distribution, it is termed as log-normal survival model. The standard log-logistic survival function is equal to

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

and the corresponding hazard function is equal to

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

Similarly the log-normal model is assumed to have bell-shaped symmetrical distribution (51) for the error term. If we assume errors to be normally distributed then the corresponding cumulative errors are also normal. The survival function of the log-normal is given by

$$S(t) = 1 - \Phi\left[\frac{lnT - \ln(\lambda)}{\sigma}\right]$$

In general, log-logistic and log-normal models are very similar and will produce similar results like logit and probit models in the regression analysis. Also, log-logistic models with p>1 and log-normal models with all possible values of the p will first rise and then fall over time.

4.5.6.1 Fitting Log-Normal and Log-Logistic distribution

The survival and fitted survival functions for lognormal parametric model are given by equations below. Table 4.7 has the analysis of the maximum likelihood estimation of parameters of the lognormal model for breast cancer patients. $S(t) = \Phi[b_0 + b_1X_1 + b_2X_2 + \cdots - k\log(t)]$; Here Φ is the cumulative distribution function of standard normal distribution.

$$\begin{aligned} S(t) = \Phi[&4.69 - 0.03X_1 + 0.27X_{21} + 0.12X_{22} - 0.2X_{23} - 0.24X_{24} - 0.82X_3 + 1.06X_{41} \\ &+ 0.76X_{42} + 0.35X_{43} + 1.96X_{51} + 1.54X_{52} + 0.8X_{53} - 0.13X_{61} - 0.38X_{62} \\ &- 0.0004X_7 - 1.07\log(t)] \end{aligned}$$

The survival and fitted survival functions for log-logistic parametric model are given by equations below. Table 4.8 has the analysis of the maximum likelihood estimation of parameters of the log-logistic model for breast cancer patients.

$$S(t;\boldsymbol{X}) = \{1 + t^k * exp(-b_0 - b_1X_1 - b_2X_2 \ldots - b_kX_k)\}^{-1}$$

$$S(t;\boldsymbol{X}) = \{1 + t^{0.7}\exp(-4.29 + 0.03X_1 - 0.25X_{21} - 0.11X_{22} - 0.22X_{23} + 0.25X_{24} + 0.09X_3 - 1.14X_{41} - 0.84X_{42} - 0.42X_{43} - 2.06X_{51} - 1.65X_{52} - 0.86X_{53} + 0.13X_{61} + 0.39X_{62} + 0.0004X_7)\}^{-1}$$

Table 4.8 Analysis of MLEs for Log-Normal Distribution

Analysis of Maximum Likelihood Parameter Estimates					
Parameter	**Estimate**	**S. E.**	**95% Confidence Limits**		**Pr > ChiSq**
Intercept	4.1885	0.0877	4.0167	4.3604	<.0001
AGE	-0.0277	0.0006	-0.0289	-0.0265	<.0001
GRADE	0.2947	0.0343	0.2274	0.3619	<.0001
GRADE	0.1404	0.0299	0.0818	0.1990	<.0001
GRADE	-0.2083	0.0300	-0.2670	-0.1495	<.0001
GRADE	-0.2638	0.0565	-0.3746	-0.1531	<.0001
GRADE	0.0000	.	.	.	.
NUMPRIMS	-0.0929	0.0139	-0.1201	-0.0656	<.0001
TREATMENT	1.2631	0.0580	1.1495	1.3768	<.0001
TREATMENT	0.9212	0.0574	0.8088	1.0337	<.0001
TREATMENT	0.4702	0.1060	0.2624	0.6780	<.0001
TREATMENT	0.0000	.	.	.	.
STAGE	2.1276	0.0447	2.0400	2.2151	<.0001
STAGE	1.6874	0.0438	1.6015	1.7733	<.0001
STAGE	0.8968	0.0493	0.8002	0.9933	<.0001
STAGE	0.0000	.	.	.	.
RACE	-0.1385	0.0243	-0.1861	-0.0909	<.0001
RACE	-0.3934	0.0325	-0.4571	-0.3298	<.0001
RACE	0.0000	.	.	.	.
TUMOR_SIZE	-0.0004	0.0001	-0.0006	-0.0002	<.0001
Scale	1.3005	0.0075	1.2859	1.3154	

Table 4.9 Analysis of MLEs for Log-Logistic Distribution

Analysis of Maximum Likelihood Parameter Estimates					
Parameter	**Estimates**	**S. E.**	**95% Confidence Limits**		**Pr > ChiSq**
Intercept	4.2928	0.0837	4.1287	4.4568	<.0001
AGE	-0.0277	0.0006	-0.0288	-0.0266	<.0001
GRADE	0.2501	0.0316	0.1881	0.3121	<.0001
GRADE	0.1071	0.0274	0.0534	0.1607	<.0001
GRADE	-0.2197	0.0275	-0.2736	-0.1658	<.0001
GRADE	-0.2483	0.0529	-0.3519	-0.1446	<.0001
GRADE	0.0000	.	.	.	.
NUMPRIMS	-0.0863	0.0127	-0.1112	-0.0614	<.0001
TREATMENT	1.1434	0.0567	1.0323	1.2545	<.0001
TREATMENT	0.8498	0.0562	0.7397	0.9599	<.0001
TREATMENT	0.4233	0.1025	0.2224	0.6242	<.0001
TREATMENT	0.0000	.	.	.	.
STAGE	2.0563	0.0426	1.9727	2.1398	<.0001
STAGE	1.6454	0.0418	1.5635	1.7274	<.0001
STAGE	0.8608	0.0467	0.7692	0.9524	<.0001
STAGE	0.0000	.	.	.	.
RACE	-0.1313	0.0231	-0.1767	-0.0860	<.0001
RACE	-0.3862	0.0305	-0.4460	-0.3263	<.0001
RACE	0.0000	.	.	.	.
TUMOR_SIZE	-0.0004	0.0001	-0.0006	-0.0002	<.0001
Scale	0.6987	0.0045	0.6899	0.7076	

4.5.6.2 Lognormal and Log-Logistic Residual Plots

To evaluate the goodness of fit for the lognormal and log-logistic models we performed a residual analysis for observed and fitted data. The result shows that the mean residual for lognormal is 0.3740, with a standard deviation of 0.3737 and range of 5.258. Log-logistic distribution has a mean residual of 0.3770, with a standard deviation of 0.3357 and a range of 4.608. Residual graphs of survival functions for lognormal and log-logistic parametric models are shown in Figure 4.5 and Figure 4.6 respectively.

Clearly the lognormal is slightly parabolic and the points does not fall close to the straight line which explains that lognormal is not the best fit for this data. From Figure 4.6 below, the graphical check of residual analysis for the log-logistic model, the graph is almost linear and hence is the winner parametric model among all others.

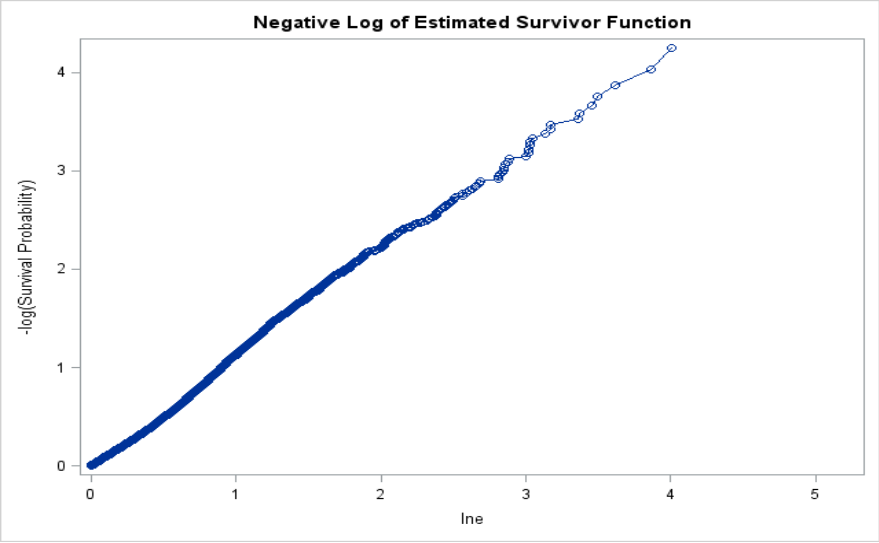


Figure 4.5 Residual plot for log-normal distribution

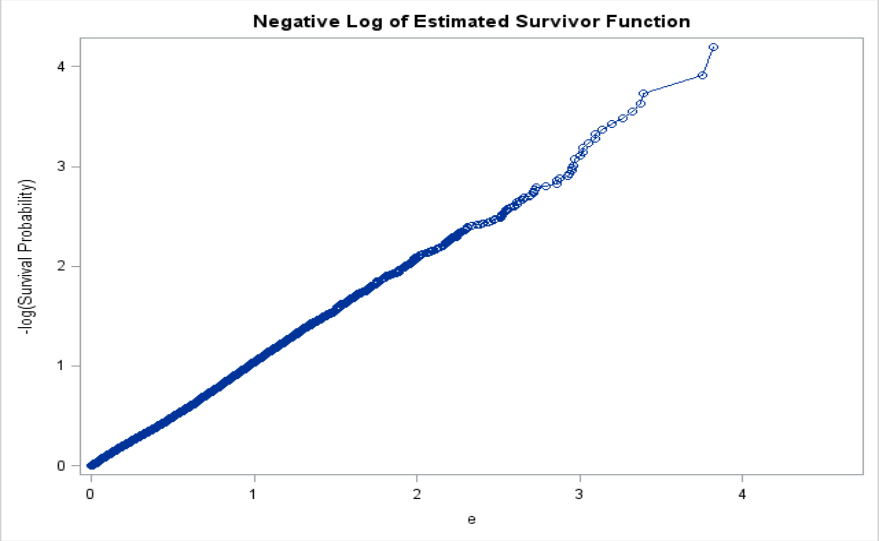


Figure 4.6 Residual plot for log-logistic distribution

4.5.7 Generalized Gamma Distribution

The survival function of the gamma distribution is the nested form of number of other distributions which is given by the equation below. Note that this model changes to log-normal as $p \rightarrow \infty$; Weibull when $k\ equals\ to\ 1$; Exponential when $k\ and\ \sigma\ equals\ to\ 1$; regular gamma distribution if $p\ equals\ to\ 1$. The main disadvantage of this generalized gamma distribution is slow and difficult to converge.

$$S(t) = 1 - \Gamma\left\{k, kexp\left[\frac{\frac{lnT_i - \lambda}{\sigma}}{p^{0.5}}\right]\right\}$$

4.5.7.1 Fitting Gamma Distribution

The survival functions for Gamma parametric model are given by equations below. Table 4.10 has the analysis of the maximum likelihood estimation of parameters of the gamma model

for breast cancer patients. Due to complexity we haven't given the fitted gamma survival function. The residual plot given in Figure 4.7, the data does not fall close to a straight line, so we conclude that gamma is not a best fit parametric model.

$$S(t;\boldsymbol{X}) = 1 - \phi_k(\lambda t) \quad \text{Where } \phi_k(\lambda t) = \int_0^x \left(\frac{(\lambda^{k-1}e^{-x})}{\Gamma(k)}\right)$$

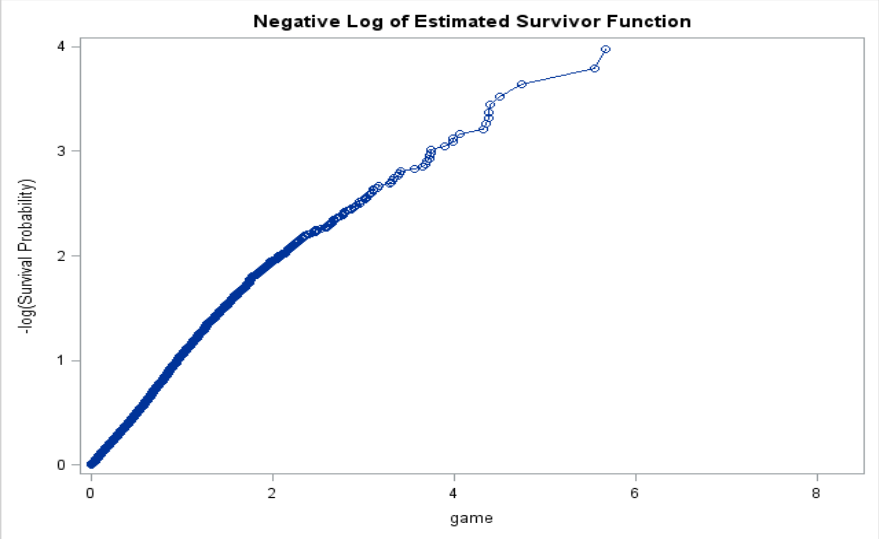


Figure 4.7 Residual plot for gamma distribution

4.5.8 Selection of best fit parametric model

We use the model selection criteria discussed in section 4.4.1 to select the best parametric model. From the previous sections, by performing the residual analysis for the fitted parametric models log-logistic parametric model performed better than other models. Also from the Table 4.9, we identify that the log-logistic model has the lowest AIC and highest likelihood values performs better than other models. This supports our choice of log-logistic model selection. From Table 4.9, we see that Gamma model is also performing close to log-logistic. To address this concern, we computed the likelihood ratio test statistic to compare these models. The test

statistic value as 31.21 and the corresponding p-value is 0.0001 which concludes that log-logistic is better. Comparison of maximum likelihood estimates for all parametric models is given in Table 4.10.

Based on the estimates of log-logistic model provided in the Table 4.10, when compared to women treated with surgery, those who received no treatment has better survival estimates compared with radiation followed by combination of radiation and surgery. However, from this model, tumor size, marital status, race has no much effect on the breast cancer.

Table 4.10 Goodness of fit for parametric models

Distribution	Log-Likelihood	AIC
Gamma	-43730.415	87506.83
Log-Normal	-43957.83687	87959.67
Weibull	-43961.92163	87967.84
Exponential	-44259.26034	88560.52
Log-Logistic	**-43714.80892**	**87473.62**

Table 4.11 Summary of MLE results for fitted parametric models

Parameter		DF	Gamma	Log-Normal	Weibull	Exponential	Log-Logistic
Intercept		1	4.6896	4.1885	5.2216	5.3589	4.2928
Age		1	-0.0284	-0.0277	-0.0287	-0.0321	-0.0277
Grade	**1**	1	0.2747	0.2947	0.2513	0.3171	0.2501
Grade	**2**	1	0.1206	0.1404	0.0974	0.1368	0.1071
Grade	**3**	1	-0.1972	-0.2083	-0.1761	-0.1723	-0.2197
Grade	**4**	1	-0.2355	-0.2638	-0.2086	-0.2108	-0.2483
Grade	**9**	0	0.0000	0.0000	0.0000	0.0000	0.0000
Numprims		1	-0.0815	-0.0929	-0.0694	-0.0851	-0.0863
Treatment	**1**	1	1.0637	1.2631	0.8678	0.9496	1.1434
Treatment	**2**	1	0.7573	0.9212	0.5897	0.6257	0.8498

Table 4.12 (Continued) Summary of MLE results for fitted parametric models

Parameter		DF	Gamma	Log-Normal	Weibull	Exponential	Log-Logistic
Treatment	**3**	1	0.3489	0.4702	0.2360	0.2614	0.4233
Treatment	**4**	0	0.0000	0.0000	0.0000	0.0000	0.0000
Stage	**1**	1	1.9563	2.1276	1.6954	1.8855	2.0563
Stage	**2**	1	1.5448	1.6874	1.3236	1.4624	1.6454
Stage	**3**	1	0.7996	0.8968	0.6660	0.7280	0.8608
Stage	**4**	0	0.0000	0.0000	0.0000	0.0000	0.0000
Race	**1**	1	-0.1340	-0.1385	-0.1255	-0.1512	-0.1313
Race	**2**	1	-0.3790	-0.3934	-0.3465	-0.3980	-0.3862
Race	**3**	0	0.0000	0.0000	0.0000	0.0000	0.0000
Tumor size		1	-0.0004	-0.0004	-0.0003	-0.0004	-0.0004
Scale		1	1.0748	1.3005	0.8484	1.0000	0.6987

4.6 Semi Parametric Analysis: Cox PH regression

The main disadvantage of non-parametric analysis is that it can only compare the survival functions of a limited number of groups whereas the parametric analysis has disadvantage of two assumptions as discussed in previous section. There is an intermediate technique whereby only an assumption is made about the way that the explanatory variables. This technique is called semi-parametric analysis, or Cox-regression. Proportional hazards regression (56) assumes that different groups have proportional hazard functions. Suppose with two groups A and B, there is a common hazard function $h(t)$, which applies to group A. Being in group B multiplies the hazard by r. i.e. $h_s(t) = r.h_A(t)$

Proportional hazards regression estimates r without estimating $h(t)$. Since hazards are chances, this means that the ratio of the hazard functions can be interpreted as a relative risk or relative rate.

$$r = \frac{h_s(t)}{h_A(t)}$$

This relative risk type ratio is very desirable in explaining the risk of events for certain categories of covariates or variable of interest.

4.6.1 Assumptions underlying Proportional Hazard Modeling

1. There exists a baseline hazard function $h_o(t)$common to all individuals in all the study groups. The baseline hazard function captures the shape of the hazard function.
2. When there is a covariate (dichotomous variable) the hazard function becomes the exponential of the parameter of interest which is termed as the exponential distribution under PH modeling.
3. Another attractive feature of Cox regression is not assuming the distributions as in the case of parametric regression. Instead refers to the fact that the hazard functions are multiplicatively related.
4. Explanatory variables act only on the r not on the baseline hazard.

4.6.2 Proportional Hazard Modeling

The formulation of Cox's regression model assumes the hazard of the subject i at the time t of the form

$$h_1(t) = h_0(t)\exp(X_i\beta)$$

Given two covariate profiles(Z_1,Z_2) the hazard ratio $\frac{h(t|z_1)}{h(t|z_2)} = exp(\frac{(Z_1-Z_2)}{\beta})$ is constant in time. Usually β is of the main interest and can be estimated independently by the partial likelihood approach (57) when right-censored data are observed. This appealing property of the PH model, together with its great flexibility, has made it one of the most popular models in

survival analysis during the past three decades. For the two-sample semi-parametric modeling, the proportional hazards model is perhaps the most widely used model and under this model, the hazard ratio for the two groups is a constant. Sometimes the constant hazard ratio may be in questioned and in this case, one can use the proportional odds model, which allows the time-dependent hazard ratio. One shortcoming of these models is that they do not apply if the two hazard or survival functions cross and this can happen in, for example, a medical study where a Treatment may be effective in long run but can have certain adverse effects during the early stage. For investigating whether there is really a difference between the two groups or whether there is really a treatment effect (58), we test the null hypothesis H_0: $\beta_1 = 0$ against the alternative hypothesisH_1: $\beta_1 \neq 0$. One has to take $T = \frac{\widehat{\beta}_1}{S(\widehat{\beta}_1)}$ as testing statistic with $\hat{\beta}_1$ being the estimate of β_1 and $S(\hat{\beta}_1)$ being the corresponding standard error. The distribution of the testing statistic is approximated by the standard normal distribution under the null hypothesis. The null hypothesis is rejected if $T \leq -c$ or $\geq c$ .The advantage is that the results can no longer be influenced by assumptions about time-dependence, since no such assumptions are made. The disadvantages are that hypotheses about time dependence can no longer be tested and that parametric analysis yields more precise estimates than the semi-parametric analysis if the assumptions about the time dependence are correct.

4.6.3 Cox Proportional Hazards Regression for breast cancer data

Using the same breast cancer survival data used in parametric survival analysis, in this section, we will examine cox regression models for the hazard function $h(t)$. As with other regression models, the identification of significant covariates and the interpretation of the estimated model coefficients is of primary interest. We will identify the likelihood that an

individual alive at time t (with the specific set of covariates as described in parametric survival modeling section) will experience the event of interest in the next very small time period. The Cox proportional hazard model (58) is used to determine the difference of survival time between races, age at diagnosis, stage of cancer, treatment, tumor size, grade, marital status and number of primary tumors. The variables in the model are introduced stepwise. The fitted Cox model reached its convergence. The model fit statistics are given below in Table 4.11. The results of three tests (likelihood, score and Wald tests) given below in Table 4.12 are used to test the hypothesis of whether the full model with all variables is better than no variables in the model. The p-value for all the three tests supported the model with all variables is statistically significant. The parameter estimate values of semi parametric cox regression model along with hazard ratios are given in Table 4.14.

From the Table 4.14, we can say that every year of age hazard increases by 3%. White women have 16% and African women has 50% greater hazard than other race women. When compared to women who are treated with surgery, those who are treated with radiation has 50.5% and women who did not receive any treatment has 64% lower hazard rate. While the combination of both surgery and radiation has 24% lower hazard rate. Type 3 tests are used to test whether there are any differences in event rate across any of the levels of the covariates used in the model. P-values reported in Table 4.13 indicate that there are significant differences in mortality between the levels of covariates. The fitted Cox PH Survival and Hazard equations for breast cancer patients are:

$$h_i(t) = h_0(t)exp(0.033X_1 - 0.31X_{21} - 0.13X_{22} + 0.19X_{23} + 0.23X_{24} + 0.082X_3 - 1.03X_{41} - 0.70X_{42} - 0.26X_{43} - 2.01X_{51} - 1.57X_{52} - 0.78X_{53} + 0.15X_{61} + 0.41X_{62} + 0.0003X_7)$$

$$S_i(t) = \exp\left(-\int_0^t \{h_0(t)exp(0.033X_1 - 0.31X_{21} - 0.13X_{22} + 0.19X_{23} + 0.23X_{24} + 0.082X_3 - 1.03X_{41} - 0.70X_{42} - 0.26X_{43} - 2.01X_{51} - 1.57X_{52} - 0.78X_{53} + 0.15X_{61} + 0.41X_{62} + 0.0003X_7)\}\, du\right)$$

Table 4.13 Cox regression model fit statistics

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
-2 LOG L	368614.59	358231.90
AIC	368614.59	358269.90
SBC	368614.59	358417.91

Table 4.14 Test results for beta coefficients

Testing Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	10382.6831	19	<.0001
Score	14594.0408	19	<.0001
Wald	12414.3302	19	<.0001

Table 4.15 Type III tests for levels of covariates

Type 3 Tests			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Age	1	2677.5352	<.0001
M_status	5	219.4995	<.0001
Grade	4	543.3288	<.0001
Race	2	161.6861	<.0001
Treatment	3	677.9500	<.0001
Stage	3	3555.2608	<.0001
Numprims	1	36.2624	<.0001
Tumor size	1	21.5580	<.0001

Table 4.16 Cox parameter estimates and hazard ratios

Parameter		DF	Parameter Estimate	Standard Error	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Age		1	0.03328	0.00064	1.034	1.033	1.035
Grade	**1**	1	-0.31020	0.03505	0.734	0.685	0.786
Grade	**2**	1	-0.12712	0.02937	0.880	0.831	0.932
Grade	**3**	1	0.19351	0.02919	1.212	1.145	1.284
Grade	**4**	1	0.23365	0.05602	1.259	1.128	1.406
Race	**1**	1	0.15031	0.02654	1.163	1.104	1.225
Race	**2**	1	0.40886	0.03375	1.504	1.408	1.607
Treatment	**1**	1	-1.03141	0.05072	0.363	0.329	0.401
Treatment	**2**	1	-0.70261	0.04975	0.505	0.458	0.557
Treatment	**3**	1	-0.25560	0.08686	0.763	0.644	0.905
Stage	**1**	1	-2.01361	0.03759	0.138	0.128	0.148
Stage	**2**	1	-1.57326	0.03686	0.213	0.198	0.229
Stage	**3**	1	-0.78178	0.04177	0.459	0.423	0.498
Numprims		1	0.08168	0.01353	1.085	1.056	1.114
Tumor size		1	0.000384	0.0000826	1.000	1.000	1.001

Finally we obtained the Cox PH survival function model for each of the three races respectively. The fit equations are given below.

Cox PH Survival and Hazard equations for White woman

$$h_i(t) = h_0(t)exp(0.04X_1 - 0.32X_{21} - 0.11X_{22} + 0.22X_{23} + 0.28X_{24} + 0.08X_3 - 0.94X_{41} - 0.56X_{42} - 0.17X_{43} - 1.97X_{51} - 1.54X_{52} - 0.81X_{53} + 0.0005X_7)$$

$$S_i(t) = \exp\left(-\int_0^t \{h_0(t)\exp(0.04X_1 - 0.32X_{21} - 0.11X_{22} + 0.22X_{23} + 0.28X_{24} + 0.08X_3 - 0.94X_{41} - 0.56X_{42} - 0.17X_{43} - 1.97X_{51} - 1.54X_{52} - 0.81X_{53} + 0.0005X_7)\}\,du\right)$$

Cox PH Survival and Hazard equations for African American woman

$$h_i(t) = h_0(t)exp(0.02X_1 - 0.28X_{21} - 0.25X_{22} + 0.03X_{23} + 0.11X_{24} + 0.08X_3 - 1.22X_{41} - 1.02X_{42} - 0.72X_{43} - 2.07X_{51} - 1.62X_{52} - 0.75X_{53} + 0.00007X_7)$$

$$S_i(t) = \exp\left(-\int_0^t \{h_0(t)exp(0.02X_1 - 0.28X_{21} - 0.25X_{22} + 0.03X_{23} + 0.11X_{24} + 0.08X_3 - 1.22X_{41} - 1.02X_{42} - 0.72X_{43} - 2.07X_{51} - 1.62X_{52} - 0.75X_{53} + 0.00007X_7)\}\,du\right)$$

Cox PH Survival and Hazard equations for other race woman

$$h_i(t) = h_0(t)exp(0.02X_1 - 0.18X_{21} - 0.11X_{22} + 0.2X_{23} - 0.08X_{24} + 0.08X_3 - 1.03X_{41} - 0.95X_{42} - 0.03X_{43} - 2.12X_{51} - 1.59X_{52} - 0.72X_{53} + 0.0003X_7)$$

$$S_i(t) = \exp\left(-\int_0^t \{h_0(t)exp(0.02X_1 - 0.18X_{21} - 0.11X_{22} + 0.2X_{23} - 0.08X_{24} + 0.08X_3 - 1.03X_{41} - 0.95X_{42} - 0.03X_{43} - 2.12X_{51} - 1.59X_{52} - 0.72X_{53} + 0.0003X_7)\}\,du\right)$$

Table 4.17 Estimates of Cox and Log-logistic models

Parameter		**DF**	**Cox**		**Log-Logistic**	
			Estimates	**Hazard**	**Estimates**	**Hazard**
Age		1	0.03328	1.034	-0.0277	1.028087
Grade	**1**	1	-0.31020	0.734	0.2501	1.284154
Grade	**2**	1	-0.12712	0.880	0.1071	1.113046
Grade	**3**	1	0.19351	1.212	-0.2197	1.245703
Grade	**4**	1	0.23365	1.259	-0.2483	1.281844
Race	**1**	1	0.15031	1.163	-0.1313	1.14031
Race	**2**	1	0.40886	1.504	-0.3862	1.471379
Treatment	**1**	1	-1.03141	0.363	1.1434	3.137417
Treatment	**2**	1	-0.70261	0.505	0.8498	2.339179
Treatment	**3**	1	-0.25560	0.763	0.4233	1.526992
Stage	**1**	1	-2.01361	0.138	2.0563	7.816993
Stage	**2**	1	-1.57326	0.213	1.6454	5.183083
Stage	**3**	1	-0.78178	0.459	0.8608	2.365052
Numprims		1	0.08168	1.085	-0.0863	1.090133
Tumor size			0.000384	1.000	-0.0004	1.000384
AIC			358269.90		87473.62	
Log likelihood			-179115.95		-43714.81	

4.7 Comparison of Survival Curves

The Table 2.16 below has the details about all the fit models with and without covariates. Log-logistic model outperformed Cox. However based on the data and attributable variables available, one can choose their best model. Table 2.15 has the comparison of Log-logistic and Cox PH estimates along with the hazard ratios.

Table 4.18 Comparison of Parametric and Cox PH models

Models	**Without Covariates**			**With Covariates**		
Distribution	**Parameters**	**-Log Likelihood**	**AIC**	**Parameters**	**-Log Likelihood**	**AIC**
Gamma	3	49170.0137	98346.03	16	43730.415	87506.83
Log-Normal	2	49366.8675	98737.74	15	43957.837	87959.67
Weibull	2	49244.9480	98493.90	15	43961.9216	87967.84
Exponential	1	49280.6023	98563.20	14	44259.2603	88560.52
Log-Logistic	2	49175.7691	98355.54	15	43714.8089	87473.62
Cox PH	-	184307.293	368614.59	16	179115.950	358269.90

4.8 Conclusion

Women who are treated with radiation alone have a median survival of 154 months. And women treated with surgery alone and both radiation & surgery reported a median survival of 25 months. Non-parametric method for survival, based on the treatment indicated that the combination of radiation and surgery has the same effect on survival as treated with surgery alone. Also from the results of Table 4.2, women in stage-4 breast cancer can be advised to stay away from any treatment for a better survival. Financially, this could really save so much for women. Further we investigated the effect of treatment stage wise. It is an interesting observation that women who are identified with malignant breast cancer tumor, but have not received any

treatment has more survival rate when compared to women who are treated with either radiation or surgery or combination of both.

This result is also supported by the results in Table 4.2. After analyzing the breast cancer data using the non-parametric Kaplan Meier method, we further performed a multivariate approach parametrically and semi-parametrically. In parametric survival modeling, we modeled the data using exponential, Weibull, log-normal, log-logistic and generalized gamma. Based on the AIC and log-likelihood comparison, log-logistic resulted as the best fit model for the data. Residual plots for the log-logistic model also fall close to the straight line, supporting our choice of parametric model.

Both intercepts and beta coefficients for almost all variables except for the women who are singled, widowed and separated, in the model are significantly differ from 0 at 0.05 level. Finally, we modeled Cox semi-parametric regression model and tabulated the hazard results with 95% confidence intervals. Neither parametric nor the Cox semi-parametric models provided any evidence about significant differences in covariates stage, race, grade and treatment. Based on AIC, as anticipated, all parametric models were performed better than the cox models.

CHAPTER FIVE

Breast Cancer Stage Classification using Multilayer Neural Networks using various Activation functions

5.1 Introduction

Artificial Neural Networks (also called connectionist models or parallel distributed processing systems) whose architecture and operation are inspired from our knowledge about biological neural cells (neurons) in the brain (59). Artificial Neural Networks (ANNs) can be described either as mathematical and computational models for non-linear function approximation, data classification, clustering and non-parametric regression or as simulations of the behavior of collections of model biological neurons. These are not real neurons in the sense that they do not model the biology, chemistry or physics of real neuron. They do, however, model several aspects of information combining and pattern recognition behavior of real neurons in a simple yet meaningful way.

Conceptually, Artificial Neutral Networks are computing constructs which mimic the process of the human brain.

Mathematically, they are a system of linked parallel equations which are solved simultaneously and iteratively (60).

Artificial Neural Networks (ANNs) or in short neural networks (NNs), like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification (61), through a learning process. Learning in biological systems involves

adjustments to the synaptic connections that exist between the neurons. This is true for ANNs as well.

The power and usefulness of ANNs have been demonstrated in several applications including speech synthesis (62), diagnostic problems and medicine (63), business and finance, robotic control (64), signal processing (65), computer vision and many other problems that fall under the category of pattern recognition.

Neural Networks has a large appeal to many researchers due to their great closeness to the structure of the brain, a unique characteristic not shared by many traditional systems.

In an analogy to the brain, an entity made up of inter connected neurons, neural networks are made up of interconnected processing elements called units (or nodes), which respond in parallel to a set of input signals given to each unit. The unit is the equivalent to its brain counterpart, the neuron.

A typical neural network consists of four main parts:

4. Processing units $\{u_j\}$, where each u_j has a certain activation level $a_j(t)$ at any point in time t.
5. Weighted interconnection between the various processing units which determine how the activation of one unit leads to input for another unit.
6. An activation rule which acts on the set of input signals at a unit to produce a new output signal, or activation.
7. Optionally, a learning rule that specifies how to adjust weights for a given input output pair.

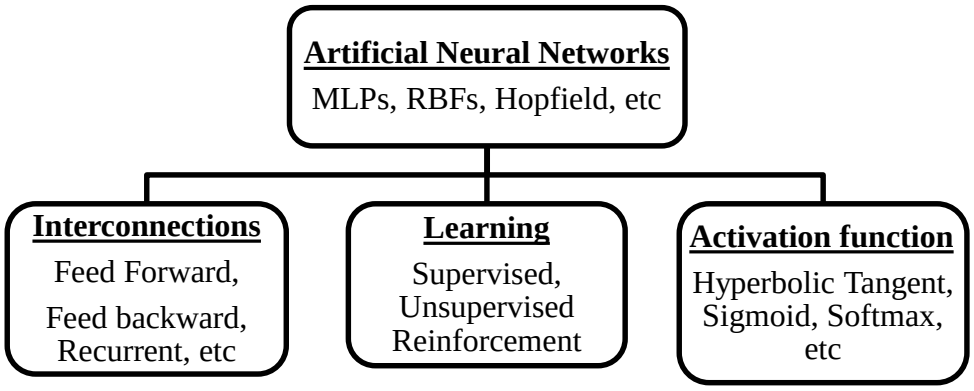


Figure 5.1 Architecture of ANN

5.1.1 Questions of Interest

Q1: Are MLP neural networks applicable to stage classification problems in Breast cancer research?

Q2: Under what conditions can MLP type neural networks be applied to stage classification problems in breast cancer data?

Q3: What are the different kind of activation functions available in MLP neural networks?

Q4: Which activation function in the training and testing of the ANN give the better performance?

Q5: What is the best activation function that can be applied to neural networks for stage classification problems in Breast cancer research?

Q6: How to evaluate the identified MLP type neural networks with different activation functions to classify breast cancer stages?

Q7: After dropping the attributable variables from the full model that contribute less in breast cancer stage classification, does the reduced model perform the same as the full model?

5.2 The First Step: McCulloch-Pitts Model

Using one of the characteristics of the biological neuron, McCulloch and Pitts (66) proposed a model for artificial neuron. The neuron model proposed by them is given in the Figure 5.2 below and is the one that widely used in ANNs with some minor modifications on it.

The artificial neuron given in the Figure 5.2 has N inputs, denoted as $u_1, u_2, \ldots, u_n$. Each line connecting these inputs to the neuron is assigned a weight, which are denoted as $w_1, w_2 \ldots, w_n$ respectively. Weights in the artificial neuron corresponding to the synaptic connections in biological neurons. The threshold in artificial neuron is usually represented by θ and the activation corresponding to the graded potential is given by the formula:

$$a = \left(\sum_{i=1}^{N} u_i w_i\right) + \theta$$

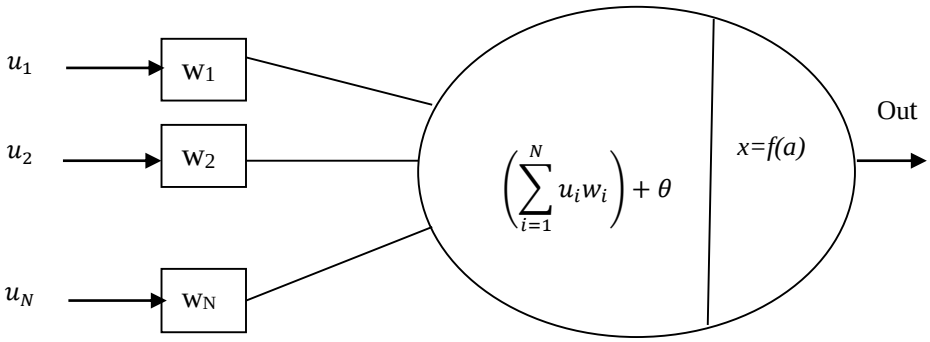


Figure 5.2 Mc Culloch-Pitts Model

5.3 A brief history of ANNs

Neural network simulations appear to be a recent development. However, this field was established before the advent of computers, and has survived at least one major setback and several areas. Many important advances have been boosted by the use of inexpensive computer emulations. Following an initial period of enthusiasm, the field survived a period of frustration

and disrepute. During this period when funding and professional support was minimal, important advances were made by relatively few researchers. These pioneers were able to develop convincing technology which surpassed the limitations identified by Minsky and Papert. Minsky and Papert (67), published a book in 1969 in which they summed up a general feeling of frustration against neural networks among researchers, and was thus accepted by most without further analysis. Currently, the neural network field enjoys a resurgence of interest and a corresponding increase in funding.

5.4 Timeline of ANN

1943 McCulloch and Pitts (66) proposed the McCulloch-Pitts neuron model.

1949 Hebb published his book "The Organization of Behavior" in which the Hebbian learning rule was proposed.

1958 Rosenblatt introduced the simple single layer networks called Perceptrons.

1969 Minsky and Papert's (67) book "Perceptrons" demonstrated the limitation of single layer perceptrons, and almost the whole field went into hibernation.

1970's and 1980's: ANN renaissance

1982 Hopfield published a series of papers on Hopfield networks.

1982 Kohonen developed the self-Organizing Maps that now bear his name.

1986 The Back-Propagation learning algorithm for Multi-Layer Perceptrons was re-discovered and the whole field got attention.

1989 Tsividis: Implemented Neural Network on a chip

1990 The sub-field of Radial Basis Function Networks was developed.

2000 The power of Ensembles of Neural Networks and support vector Machines becomes apparent.

5.5 Inspiration for ANN: Biological Prototype

Much is still unknown about how the brain trains itself to process information, so theories abound (Figure 5.4).In the human brain, a typical neuron collects signals from others through a host of fine structures called Dendrites. The neuron sends out spikes of electrical activity through a long, thin stand known as an axon, which splits into thousands of branches. At the end of each branch, a structure called a synapse converts the activity from axon into electrical effects that inhibit or excite activity in the connected neurons. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory inputs, it sends a spike of electrical activity down its axon (Figure 5.3). Learning occurs by changing the effectiveness of the synapse so that the influence of one neuron on another changes.

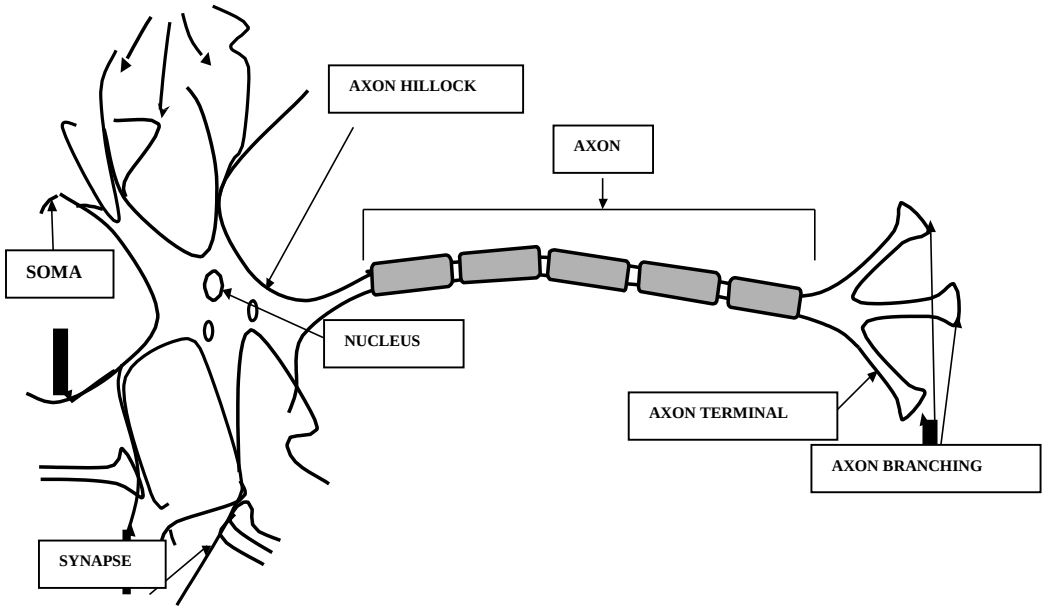


Figure 5.3 Biological Neuron

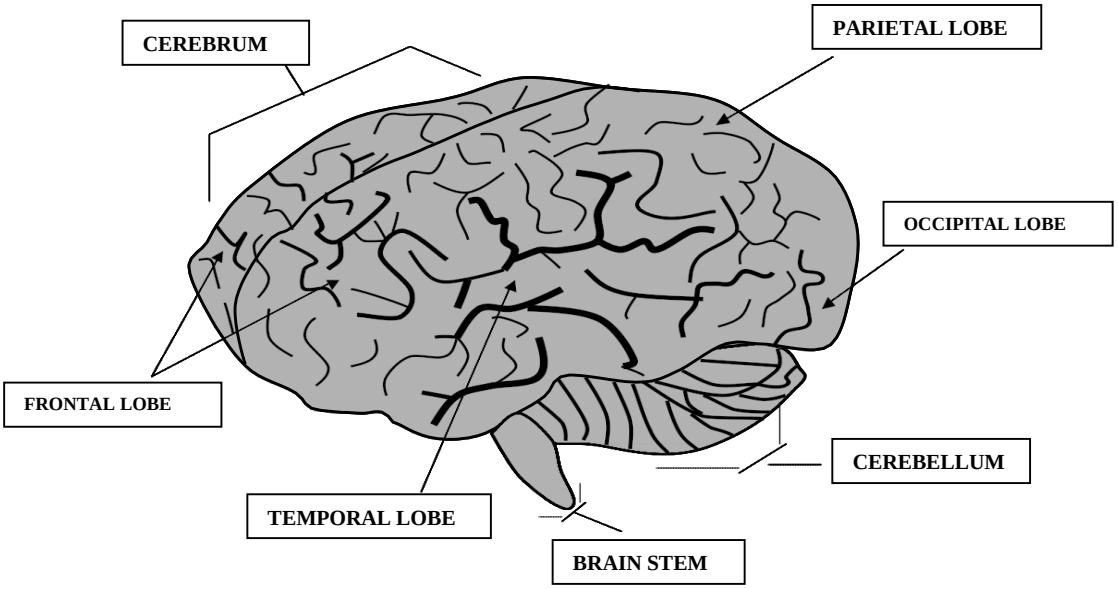


Figure 5.4 Human Brain

5.6 Brain versus Computers: Some interesting numbers

1. There are approximately 10 billion neurons in the human cortex, compared with thousands of processors in the most powerful parallel computers.
2. Each biological neuron is connected to several thousands of other neurons, similar to the connectivity in powerful parallel computers.
3. Lack of processing units can be compensated by speed. The typical operating speeds of biological neurons (68) is measured in milliseconds (10^{-3}s), while a silicon chip can operate in nanoseconds (10^{-9}s).
4. The human brain is extremely energy efficient, using approximately 10^{-6} joules per operation per second, where as the best computers today use around 10^{-16} joules per operation per second.
5. Brains have been evolving for tens of millions of years; computers have been evolving for tens of decades.

5.7 ANN Types

Feed forward: Single Layer Perceptron (69), MLP, ADALINE (Adaptive Linear Neuron) (70), RBF.

Self-Organized: SOM (Kohonen Maps).

Recurrent: Simple Recurrent Network, Hopfield Network (71).

Stochastic: Boltzmann machines (72), RBM.

Modular: Committee of Machines, Associative Neural Networks (ASNN), Ensembles.

Others: Instantaneously trained, Spiking Neural Networks (SNN) (73), Dynamic, Cascades, Neuro Fuzzy (74), PPS, GTM (75).

5.8 Learning methods in ANN

As listed in previous section, there are many forms of neural networks. Most operate by passing neural 'activations' through a network of connected neurons. One of the most powerful features of neural networks is their ability to learn and generalize from a set of training data. They adapt the strengths/ weights of the connections between neurons so that the final output activations are correct.

There are three broad types of learning:

1) Supervised learning (i.e., learning with a teacher)
2) Unsupervised learning (i.e., learning with no help)
3) Reinforcement learning (i.e., learning with limited feedback)

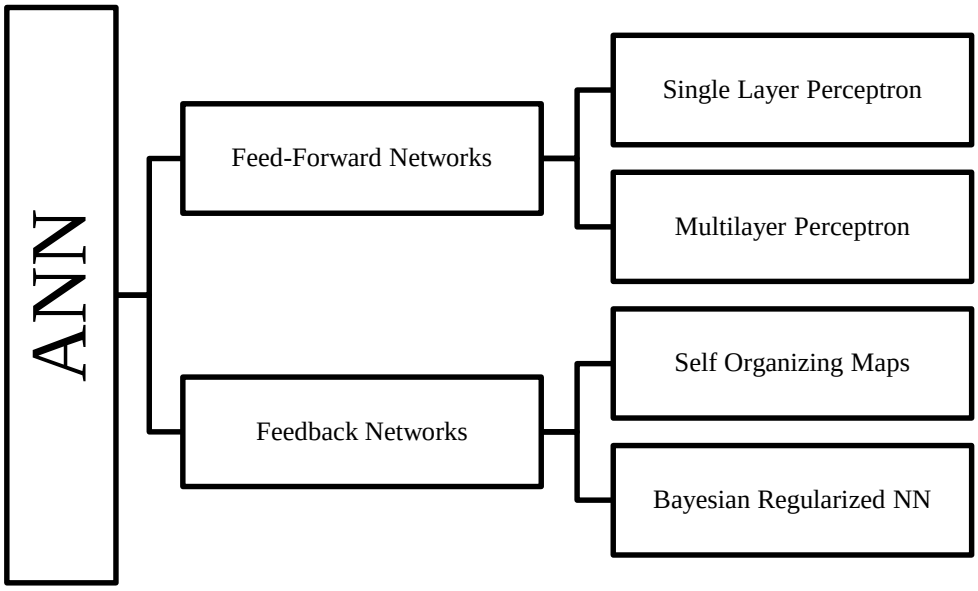


Figure 5.5 ANN Architecture

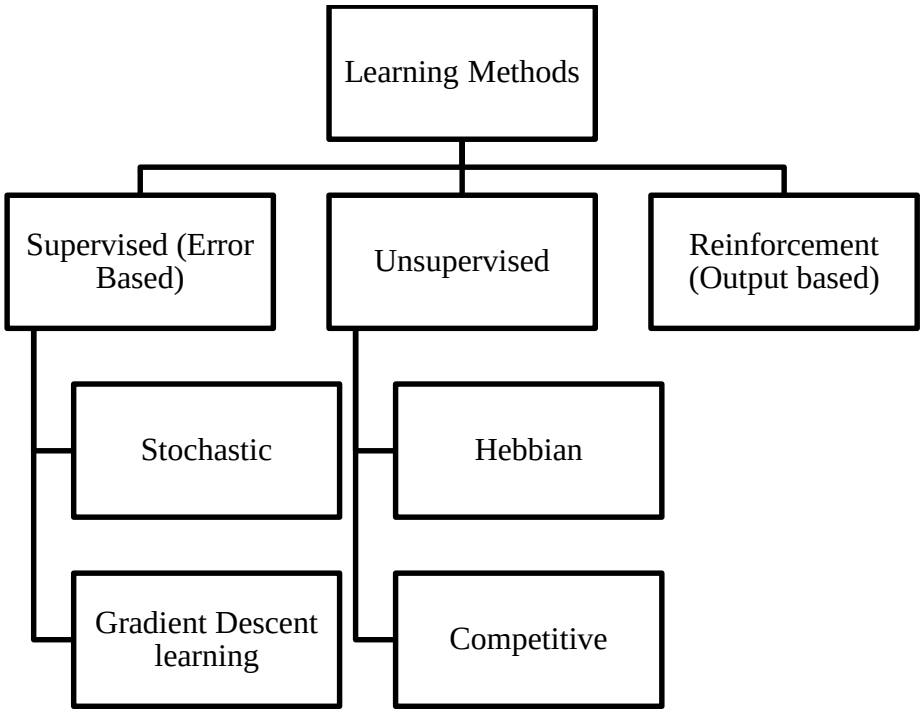


Figure 5.6 Learning Methods in ANN

5.8.1 Supervised learning

Which incorporates an external teacher, so that each output unit is told what its desired response to input signals ought to be. In this mode, the actual output of a neural network is compared to the desired output. Weights, which are usually randomly set to begin with, are then adjusted by the network so that the next iteration, or cycle, will produce a closer match between the desired and the actual output. The learning method tries to minimize the current errors of all processing elements. This global error reduction is created over time by continuously modifying the input weights until acceptable network accuracy is reached. Paradigms of supervised learning include error-correction learning reinforcement learning and stochastic learning (76).

With supervised learning, the Artificial Neural Network must be trained before it becomes useful. Training consists of presenting input and output data to the network. That is, for each input set provided to the system the corresponding desired output set is provided as well. This training is considered complete when the neural network reaches a user defined performance level.

An important issue concerning supervised learning is the problem of error convergence, i.e., the minimization of error between the desired and computed unit values. The aim is to determine a set of weights which minimizes the error. One well-known method, which is common to many learning paradigms is the Least Mean Square (LMS) convergence (77).

5.8.2 Unsupervised learning

Uses no external teacher and is based upon only local information, it is also referred to as self-organization, data presented to the network and detects their emergent collective properties. Paradigms of unsupervised learning are Hebbian learning and competitive learning. From

Human Neurons to Artificial Neuron Esther aspect of learning concerns the distinction or not of a separate phase, during which the network is trained, and a subsequent operation phase. We say that a neural network learns off-line if the learning phase and the operation phase are distinct. A neural network learns on-line if it learns and operates at the same time. Usually, supervised learning is performed off-line, whereas unsupervised learning is performed on-line.

A simple version of Hebbian learning rule (78) is that when unit i and unit j are simultaneously excited, the strength of the connection between them increases in proportion to the product of their activations.

In competitive learning, if a new pattern is determined to belong to a previously recognized cluster, then the inclusion of the new pattern into that cluster will affect the representation (e.g., centroid) of the cluster. This will in turn change the weights characterizing the classification network. If the new pattern of 'input-outputs' determined to belong to none of the previously recognized cluster, then (the structure and the weights of) the network will be adjusted to accommodate the new class (cluster).

5.8.3 Reinforcement learning

For many applications, the desired output may not be known precisely. Other learning law have been developed based on the information whether the response is correct or wrong. This mode of learning is called reinforcement learning or learning with critic.

There are many situations where the desired output for a given input is not known. Only the binary result that the output is right or wrong may be available. This output is called reinforcement signal. This signal only evaluates the output. The learning based on this evaluate signal is called reinforcement learning. Since this is evaluative and not instructive, it is also called learning with critic as opposed to learning with teacher in the supervised learning.

5.9 Multilayer Perceptron and Radial Basis Function

Multilayer perceptrons (MLPs) and radial basis function (RBF) networks are the two most commonly-used types of feed forward network. They have much more in common than most of the neural network literature would suggest. The only fundamental difference is the way in which hidden units combine values coming from preceding layers in the network--MLPs use inner products, while RBFs use Euclidean distance. There are also differences in the customary methods for training MLPs and RBF networks, although most methods for training MLPs can also be applied to RBF networks. Furthermore, there are crucial differences between two broad types of RBF network, the ordinary RBF networks and the normalized RBF networks that are ignored in most of the neural network literature. These differences have important consequences for the generalization ability of the networks, especially when the number of inputs is large. Our focus in this chapter will be on MLPs. A network with three layers: input, hidden and output layers.

An activation function $f_{(x,wi)}$connects the weights $wi$of a neuron *I* to the input *x* and determines the activation or the state of the neuron. An input function *x* of the formal neuron *I* corresponds to the incoming activity of the neuron, the weight *w* represents the effective magnitude of information transmission between neurons, the activation function $f_{(x,wi)}$ describes the main computation performed by a biological neuron and the output function out_i corresponds to the overall activity transmitted to the next neuron in the processing stream.

5.10 Activation Functions

The crucial step in MLP neural network structure is generating the net inputs by using a scalar-to-scalar function which is known as the "activation function" or "threshold function" or

"transfer function" (79). These activation functions are used to limit the amplitude of the output of a neuron. The typical activation functions which are used to solve the non-linear problems are sigmoid, tangent, softmax, radial basis functions among others. These functions further process the output of the neuron after initial processing has taken place and are non-linear in nature by transforming the weighted sum of inputs to an output value and do the final mapping. In most cases these functions squash the amplitude range to a limited value probably the normalized value. Interestingly the outputs of these functions are further processed by running more number of iterations unless the network attains the desired convergence. In back propagation learning the functions implemented should have the characteristics like the continuous, differentiable, and monotonically non-decreasing and output should be bounded.

As mentioned earlier, ANNs are mostly used in modeling nonlinear data. Neural networks because of its nonlinear structure are used either to approximate a posteriori probabilities for clustering/classification or to approximate probability densities of the training data (80, 81). Nonlinearity is introduced into an MLP network in the form of an activation function for the hidden units. The nonlinearity in the network is the reason why MLPs are so powerful. Below are few important papers surveyed which show that the choice of transfer functions is considered by some experts to be as important as the network architecture and learning algorithm.

G. Cybenko (1989), K. Hornik et al. (1989) in their research articles (82, 83) discussed about using sigmoidal functions generating sigmoidal outputs as universal approximators. However E. J. Hartman, et al. (1990) and J. Park, et al. (1991) also termed Gaussian outputs also as universal approximators (82, 83). Hartman and Keeler (1991) proposed a new activation function called Gaussian bars (84). Pao (1989) in his book “Adaptive Pattern Recognition and

Neural Networks" discussed about using a combination of various activation functions (85). Simon Haykin and Leung (1993) were very successful with using radial transfer functions (86).

Dorffner (1994) using conic section function networks introduced new transformation functions that change smoothly from sigmoidal to Gaussian-like (87). Girauld, et al. (1995) introduced simplified Gaussian functions called Lorentzian transfer functions which are widely used in many research works (88).

Two most popular feed forward neural networks models, the multi-layer perceptron (MLP) and the Radial Basis Function (RBF) networks, are based on specific architectures and the transfer functions. Below are few activation functions in detail.

5.10.1 Identity Function

The Identity function is also known a linear function. The output of the function is same as the input variable. Sometimes a constant is used to multiply it to form a linear function with scaled magnitude. The activation function needs to introduce non linearity in to the networks for the network to be robust.

$$f(x) = x$$

$f(x) = kx$ Where k is a scaling constant

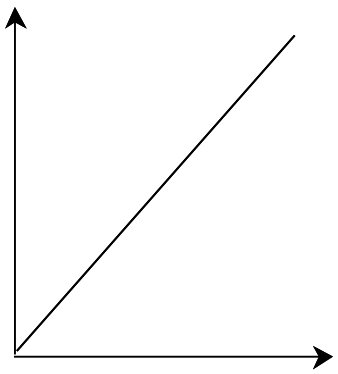


Figure 5.7 Identity Function

5.10.2 Binary Step Function

This function is also known as the Heaviside function or threshold function or hard limit function, with threshold θ. The output is always a binary value and it is decided by the function.

$$f(x) = \begin{cases} 0 \ if \ \ x \le \theta \\ 1 \ if \ \ x > \theta \end{cases}$$

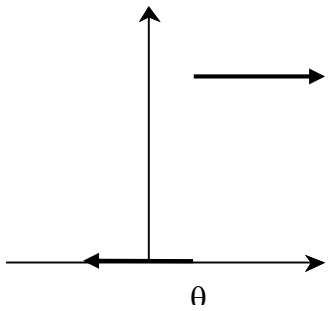


Figure 5.8 Binary Step Function

5.10.3 Saturating linear function

This function is also known as ramp function or piece wise linear sigmoid function (89) combines the Heaviside function with a linear output function.

$$f(x) = \begin{cases} 0, & x \le 0 \\ x, 0 < x < 1 \\ 1, & x \ge 1 \end{cases}$$

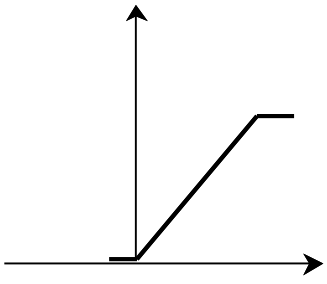


Figure 5.9 Ramp Function

5.10.4 Sigmoid Functions

Sigmoidal output functions smooth out many shallow local minima in the total output functions of the network. For classification type of problems this may be desirable, but for general mappings it limits the precision of the adaptive system (90). This is the most commonly used transfer function in MLP as it gives good results in most cases and can dramatically reduce the computation burden of training. The term sigmoid mean a graph which is 'S-shaped' curve. It is most commonly used function in the neural networks where the training is implemented by using the back propagation algorithms. The significance of this function is that the computation capacity for training is reduced and can be distinguished easily.

Uni-polar sigmoid

The output of this function is bounded to [0, 1]. The function gets zero to as the value of x tends to infinity in the negative side. Its analytic equation is given below.

$$f(x) = \frac{1}{1 + e^{x}}$$

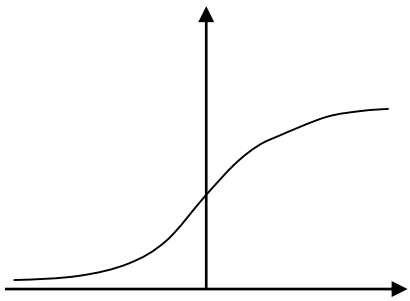


Figure 5.10 Uni-polar Sigmoid Function

Bi-Polar Sigmoid Function

The bi-polar sigmoid function is similar to the uni-polar sigmoid except that the limits of the output range between [-1, 1].

$$f(x) = \frac{1 - e^{x}}{1 + e^{x}}$$

Bipolar binary and uni-polar binary are called as hard limiting activation functions used in discrete neuron model. Uni-polar continuous and bipolar continuous are called soft limiting activation functions are called sigmoidal characteristics.

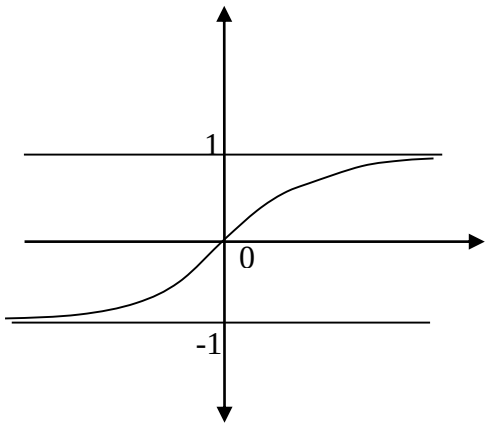


Figure 5.11 Bi-Polar Sigmoid function

5.10.5 Hyperbolic Tangent Function

The hyperbolic transfer function also ranges between [-1, 1]. This function is implemented in the replication of the sigmoid function where the output range is varying between -1 to 1.

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{\sinh x}{\cosh x} = \tanh x$$

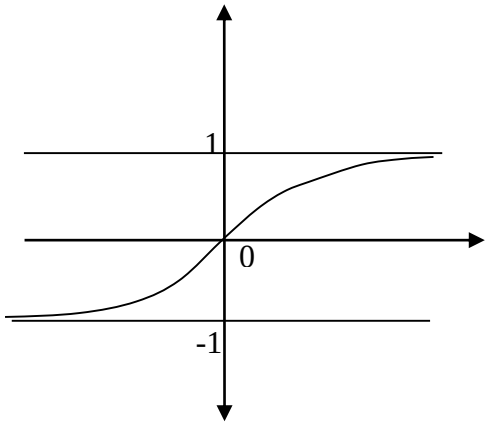


Figure 5.12 Hyperbolic Tangent function

5.10.6 Radial basis functions (RBFs)

As MLP's implement sigmoidal transfer functions, RBFs typically use Gaussian functions. Both types of networks are universal approximators. This is an important, but almost trivial property, since any network using non-polynomial transfer functions are always universal approximators. The speed of convergence and the complexity of these networks to solve a given problem is more interesting.

$$g(x, c) = g(||x - c||)$$

$$y(x) = \sum_{i=1}^{N} w_i g\left(||x - c_i||\right)$$

Where$y(x)$ is represented as a sum of N radial basis functions and each of them are associated with a different center c_i and weighted by an appropriate weight $w_i$and $w_i$can be obtained by the matrix methods of linear least squares.(91)

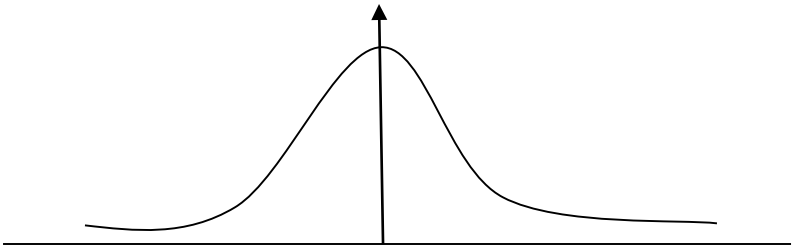


Figure 5.13 Radial basis function

5.11 Evaluation of model performance

The methods used for the model performance evaluation of different neural networks include comparison of area under ROC curves, positive predictive values (PPVs) and overall accuracy. The values of training and testing the full and reduced models were evaluated and tabulated in the following sections. In the ROC graph the diagonal line represents diagnostic test where sensitivity equals (1– specificity) which refers that the test has no diagnostic value. A test where both sensitivity and specificity are close to 1, which in turn will return a ROC value also close to 1, has good diagnostic ability.

5.11.1 Accuracy, ROC, PPVs

Receiver operating characteristic (ROC) curves (92) are frequently used to compare the diagnostic qualities of statistical models. For a given confidence threshold, the fraction of negative outcomes that are correctly identified as negatives is called the true-positive fraction (TPF = sensitivity) and the fraction of the positive outcomes that are correctly identified is called the true-negative fraction (TNF = specificity). The false-positive fraction (FPF) and the false-negative fraction (FNF) are defined in the same way. Confusion matrix generated for a model gives all these details of classification. For the actually positive and the actually negative outcomes, probability distributions can be derived for the various states of truth.

Table 5.1 Classification Table

X	Actual State	
	Positive	Negative
Considered positive	True positive (TP)	False positive (FP)
Considered negative	False negative (FN)	True negative (TN)

There are three components to predict the accuracy: the amount and quality of the data, the predictive power of the prognostic factors, and the prognostic method's ability to capture the power of the prognostic factors (93). This study is mainly focused on the area under curve (AUC). The measure of comparative accuracy is the trapezoidal approximation to the area under the receiver operating characteristic curve. The area under this curve is a nonparametric measure of discrimination. While squared error summarizes how close each individual's survival prediction is to the true outcome, the receiver operating characteristic area measures the relative goodness of the set of predictions as a whole by comparing the predicted probability of each individual with that of all pairs of individual s. This area is calculated using the predictive scores of each algorithm in order to compare their average accuracy in predicting outcome. The receiver operating characteristic area is independent of both the prior probability of each outcome and the threshold cutoff for categorization, and its computation requires only that the algorithm produce an ordinal-scaled relative predictive score. In terms of mortality, the receiver operating characteristic area estimates the probability that the algorithm will assign a higher mortality score to the patient who died than to the patient who lived. The receiver operating characteristic area varies from 0 to 1. When the prognostic score is unrelated to survival, the score is 0.5, indicating chance accuracy. The farther the scoreisfrom0.5, the better, on average, the prediction model is at predicting which of the individuals who will survive. Positive predictive values

(PPVs) refers to the chance that a positive test result will be correct, negative predictive value is concerned only with negative test results. The interesting thing about positive and negative predictive values is that they change if the prevalence of the disease changes. In fact, for any diagnostic test, the positive predictive value will fall as the prevalence of the disease falls while the negative predictive value will rise (94).

5.12 Breast Cancer stage classification using various activation functions

In traditional regression, a specific equation must be predetermined based on the data in the system in order to find a relation between the inputs to output variable. Whereas the general structure of an ANN can be applied practically on any system. Also, ANNs have been shown to outperform regression models when outliers exist in the data and a MLP neural network with an appropriate activation function in the hidden layer is always considered as a better model.

The objective of using MLP neural networks in this chapter is to be able to classify stages of breast cancer data. In order to classify the stages we hves chosen MLP network as the classifier. We designed different feed forward MLP networks with one hidden layer with different inputs. One hidden layer MLP is almost always sufficient to approximate any continuous function up to certain accuracy (95). It is proven in many situations that MLPs possess the ability to learn and give the better performance especially in the case of classification. The MLP network has to be trained before it able to perform specific task with less error. In this study we used 33152 (70%) data for training, 14015 (30%) data for testing the trained network.

Table 5.2 Activation Functions

Activation function	**Definition**
Linear	$f(x) = x$
Binary step	$f(x) = \begin{cases} 0, x \le 0 \\ 1, x > 0 \end{cases}$
Ramp function or Saturating linear	$f(x) = \begin{cases} 0, x \le 0 \\ x, 0 < x < 1 \\ 1, x > 1 \end{cases}$
Uni polar Sigmoid	$f(x) = \frac{1}{1+e^{x}}$
Bi-polar	$f(x) = \frac{1-e^{-x}}{1+e^{-x}}$
Hyperbolic tangent	$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} = \frac{\sinh(x)}{\cosh(x)}$
Radial Basis Function	$g(x,c) = g(\lvert\lvert x - c \rvert\rvert)$ $y(x) = \sum_{i=1}^{N} w_i g(\lvert\lvert x - c_i \rvert\rvert)$

In this study we compared the performance of an MLP network by using different activation functions. Every MLP network consists of an input layer, hidden layer and an output layer. For all the MLPs with different activation functions hidden nodes are selected automatically based on the requirement for training. The best number of hidden nodes required in the hidden layer depends on the number of inputs and outputs, amount of noise in the

Table 5.3 Input Variables & types

Input Variables	**Modalities**	**Details**
Tumor Size	Real	1mm – 998mm
Treatment	Categorical	1= No Treatment 2= Radiation 3=Radiation & Surgery 4= Surgery
Age	Real	21 - 102
Number of primary tumors	Real	1,2,3,4,5 = able to detect 9 = not able to be detected
Grade	Categorical	1=Well differentiated 2=Moderately differentiated 3=Poorly differentiated 4=undifferentiated 9=Cell type not determined
Marital Status	Categorical	1 = Single 2 = Married 3 = Separated 4 = Divorced 5 = Widowed 9 = Unknown
Race	Categorical	1 = Whites 2 = African Americans 3 = Other races
Duration	Real	1-203 months

targets, activation function used. Rules of thumb don't usually work. The number of hidden neurons decided upon training stage of the MLP networks. Four output neurons for four stage

classification are needed to classify the class of the target outputs. The performances of the MLP networks will be evaluated in terms of percentages for correct classification, defined as the difference between the actual and the simulated results and by ROC analysis.

The work in this chapter is divided into two parts. In the first part we designed six neural networks models using all combinations of activation functions with all the inputs including tumor size, treatment, age, number of preliminary tumors, and grade of the tumor, marital status and race of women to classify their stage of breast cancer. At the end of first part of work, our objective is to find the best combination of activation function pair that classifies the breast cancer stages, by comparing the number of hidden nodes, positive predictive values (PPVs), percent of correct classification and comparing ROCs (96). Table 5.2 has the details of input variables used in modeling the neural networks. After identifying the best activation function, in our second part of work, we tried to reduce the neural network model by eliminating the inputs which perform the least. Inputs which fall below 5% normalized importance are eliminated and the networks are rerun to check the efficiency of the model.

For the first part, fixing Hyperbolic Tangent as the activation function for hidden layer, we used softmax, hyperbolic tangent, sigmoid as the transfer functions in output layer. Later fixing sigmoid function as activation function we have used the softmax, hyperbolic tangent, sigmoid as the transfer functions in output layer. This resulted in total of 6 different models.

Results of 6 full models with the percentage of correct predictions, positive predicted values (PPVs) during training and testing along with stage wise area under curve values are given in Table 5.3 and Table 5.4. From these tables, the model with hyperbolic tangent and softmax function has a better prediction with less number of hidden nodes. Figure 5.14 gives the ROC of the selected model.

Table 5.4 Full Model stage classification probabilities

Full Model details		Positive Predictive Probabilities				
Training	**Number of hidden units**	**P(1|1)**	**P(2|2)**	**P(3|3)**	**P(4|4)**	**Overall Accuracy**
HT– Softmax	8	88.9%	75.0%	41.9%	33.8%	79.0%
HT – HT	9	91.7%	73.7%	45.3%	26.3%	79.8%
HT – Sigmoid	9	90.4%	76.4%	0%	0%	77.0%
Sigmoid – Softmax	9	89.5%	74.5%	46%	26%	79.1%
Sigmoid – HT	9	90.6%	75.1%	40.7%	1.2%	79.0%
Sigmoid – Sigmoid	9	91.7%	73.7%	50.9%	0%	79.4%
Testing						
HT– Softmax	8	88.9%	75.0%	39.3%	28.8%	78.8%
HT – HT	9	92.1%	72.7%	43.0%	26.1%	79.5%
HT – Sigmoid	9	90.9%	76.1%	0%	0%	77.6%
Sigmoid – Softmax	9	89.8%	74.5%	45.1%	23.7%	79.1%
Sigmoid – HT	9	90.8%	73.5%	43.1%	0.8%	78.5%
Sigmoid – Sigmoid	9	91.7%	72.5%	51.2%	0%	79.0%

HT-Hyperbolic Tangent

Figure 5.15 and Figure 5.16 are the performance analysis of PPVs for training and testing of the full models. From these figures and the results given in Table 5.3 and Table 5.4, though the sigmoid-softmax pair has comparatively same results like hyperbolic tangent-softmax pair, we prefer to select hyperbolic tangent-softmax pair for the following reasons. A MLP model with the best performance using less number of hidden units is considered as the best ANN representing the problem. Hyperbolic tangent-softmax model uses only 8 hidden units whereas softmax-sigmoid network uses 9 hidden units. Also since the hyperbolic tangent activation function has a derivative, it can be used with gradient descent based training methods. The hyperbolic tangent activation function is perhaps the most common activation function used for neural networks. The

hyperbolic tangent function provides similar scaling to the sigmoid activation function, however, the hyperbolic tangent activation function has a range from -1 to 1. Because of this greater numeric range the hyperbolic activation function is often used in place of the sigmoid activation function. The neural network diagram for the selected full model is given in Figure 5.17.

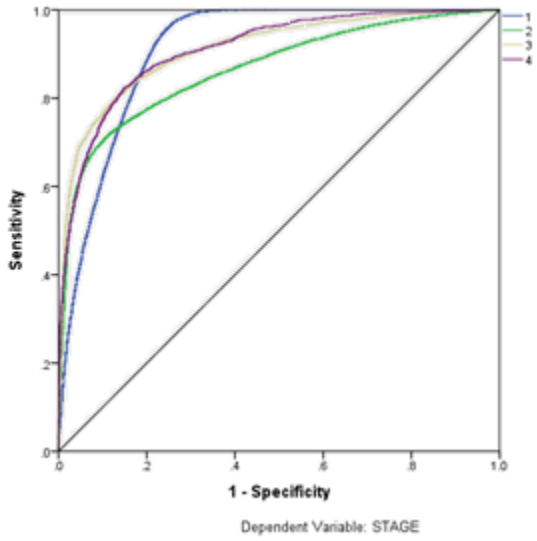


Figure 5.14 ROC of the full model

Table 5.5 ROC values of full models

Activation Functions	AUROC Stages			
	1	**2**	**3**	**4**
HT– Softmax	0.911	0.866	0.910	0.910
HT– HT	0.910	0.866	0.882	0.895
HT– Sigmoid	0.910	0.859	0.909	0.886
Sigmoid – Softmax	0.912	0.868	0.913	0.919
Sigmoid –HT	0.909	0.863	0.862	0.881
Sigmoid – Sigmoid	0.910	0.862	0.909	0.882

HT-Hyperbolic Tangent

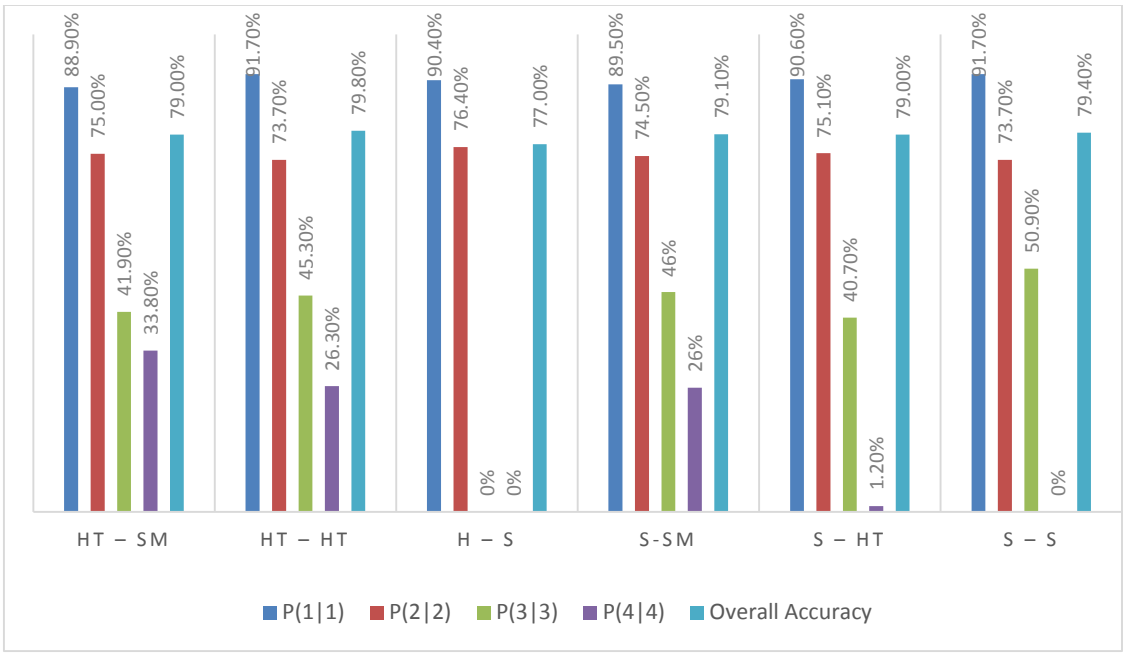


Figure 5.15 Testing performance of full models

HT-Hyperbolic Tangent; SM-Softmax; S-sigmoid

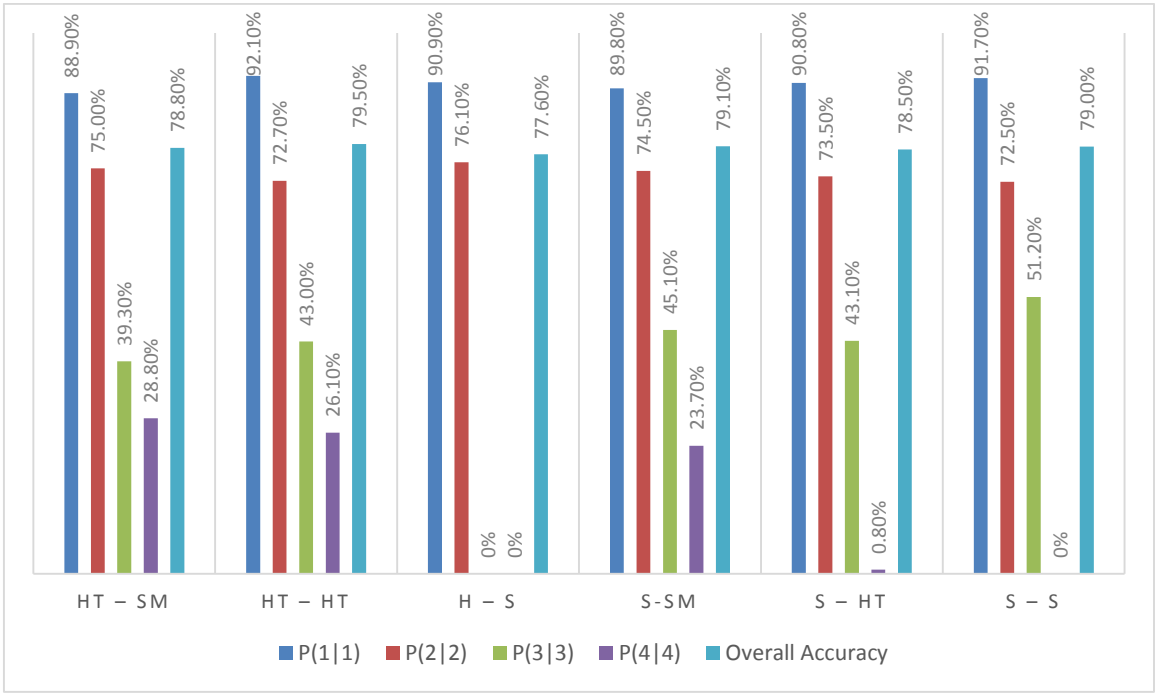


Figure 5.16 Testing performance of full models

HT-Hyperbolic Tangent; SM-Softmax; S-sigmoid

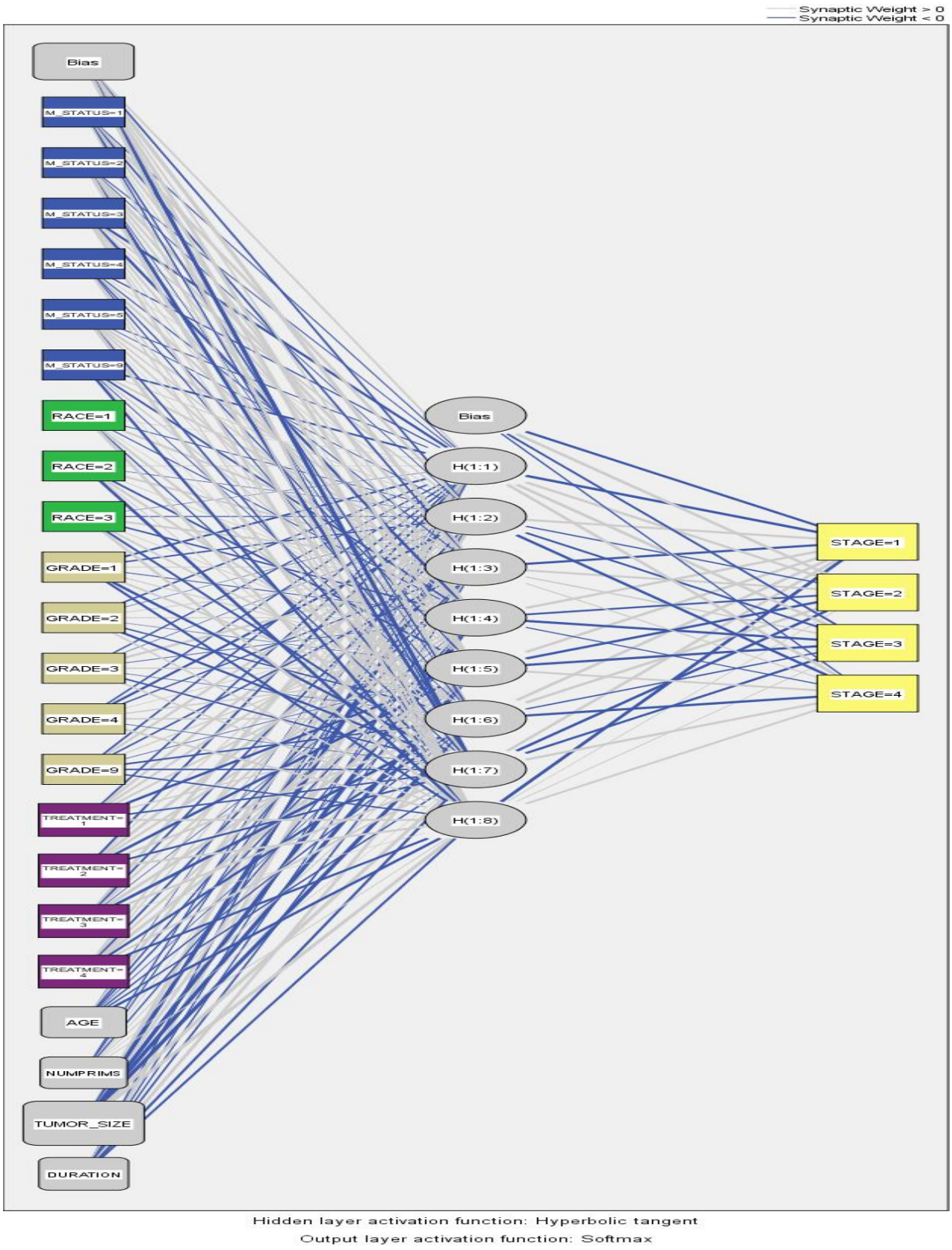


Figure 5.17 Full MLP model using Hyperbolic tangent-softmax activation function

5.13 Reduced Neural Network Model and Conclusion

After identifying that the neural network using the combination of hyperbolic tangent-softmax pair as the best neural network model for breast cancer stage classification, we further proceed to identify the reduced neural network model. Using the same activation pair selected from full model we try to find reduced model, if any, by using fewer input units and/or hidden units which can perform equivalent to full model or even better than the full model. In order to do this, we rerun a neural network model by eliminating the input variables from the full model which have less than 5% normalized importance in performance of breast cancer stage classification.

Table 5.6 Importance and Normalized Importance of input variables

Inputs	HT – SM		HT – HT		HT – S		S – SM		S – HT		S – S	
	Imp	N.Imp	Imp	N.Imp	Imp	N.Imp	Imp	N.Imp	Imp	N.Imp	Imp	N.Imp
M_STATUS	.021	**3.8%**	.064	14.6%	.090	15.3%	.025	4.2%	.037	7.6%	.040	6.9%
RACE	.018	**3.4%**	.016	3.6%	.018	3.1%	.018	3.0%	.013	2.7%	.016	2.8%
GRADE	.025	**4.6%**	.037	8.5%	.035	6.0%	.023	3.9%	.050	10.3%	.033	5.7%
TREATMENT	.127	23.4%	.150	34.3%	.069	11.7%	.111	18.7%	.134	27.6%	.110	19.3%
AGE	.059	10.9%	.106	24.3%	.095	16.2%	.048	8.1%	.103	21.3%	.118	20.7%
NUMPRIMS	.085	15.7%	.108	24.7%	.060	10.2%	.021	3.5%	.099	20.5%	.034	6.0%
TUMOR_SIZE	.543	100.0%	.437	100.0%	.586	100.0%	.593	100.0%	.484	100.0%	.573	100.0%
DURATION	.121	22.3%	.083	18.9%	.047	8.1%	.162	27.3%	.081	16.8%	.075	13.0%

From Table 5.5 for the selected activation pair of full model neural network, the input variables race, marital status and grade are the variables fall below 5% normalized importance and are eligible for elimination. Eliminating these input variables we modeled a reduced network model to perform stage classification of breast cancer. The reduced model has 8 input variables, 6 hidden units to classify breast cancer stages compared with 22 inputs and 8 hidden units of full model. An output equation of ANN will be a composite function given as

$$y_i = f\left\{\sum g\left(\sum(\bullet)\right)\right\};$$

$$where\ i = 1,2,3,4;\ f(\bullet)\ is\ the\ output\ function\ \ and\ g(\bullet) is\ a\ hidden\ layer\ outcome$$

Table 5.7 Training and Testing results of the reduced neural network model

		Positive Predictive Probabilities				
Reduced Model details	**ANN Architecture I – H – O**	**P(1|1)**	**P(2|2)**	**P(3|3)**	**P(4|4)**	**Overall Accuracy**
Training	8-6-4	89.8%	74.2%	49.8%	30.3%	79.5%
Testing	8-6-4	90.0%	73.5%	49.2%	24.9%	79.0%

I-Input units; H-Hidden units; O- Output units

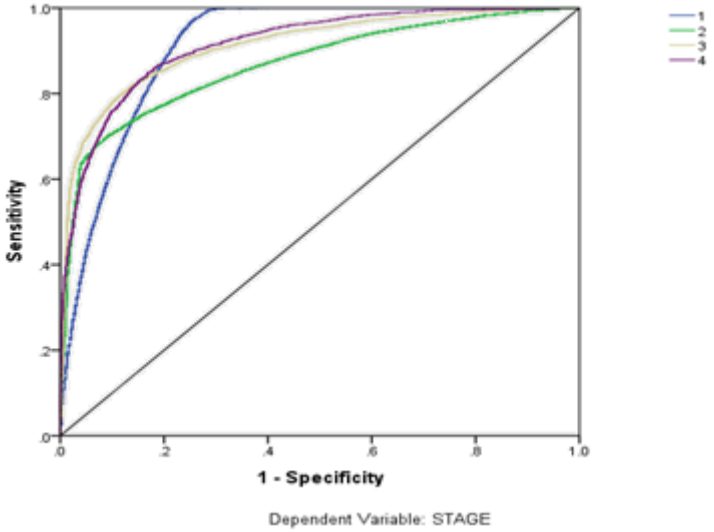


Figure 5.18 ROC of the reduced neural network model

Table 5.8 ROC Comparison for Full and reduced models

Models	**Stage-1**	**Stage-2**	**Stage-3**	**Stage-4**
Reduced Model	0.911	0.868	0.912	0.915
Full Model	0.911	0.866	0.910	0.910

Table 5.6 has the results of reduced neural network architecture, positive predictive values, and overall accuracy of training and testing classification results. The reduced model area under curve and full model area under curve results are compared and presented in Table 5.7. Reduced model works efficiently using 8 input units and 6 hidden units only. Figure 5.18 gives the ROC of the reduced model. Reduced model performed almost close to the full model but with fewer units in input and hidden layers. Clearly the reduced model with hyperbolic tangent-softmax activation pair is opted as précised one for breast cancer stage classification.

CHAPTER SIX

A Comparison of Artificial Neural Network and Decision trees with Logistic Regression as Classification Models for Breast Cancer Survival

6.1 Introduction

Computer models are being employed actively in the clinical diagnostic field to differentiate between healthy and disease suffering patients. These computer models are responsible in facilitation of making accurate decisions towards likelihood of disease based on certain characteristics of the patient. Many different modeling techniques have been developed, tested and refined. These techniques include both statistical (Linear Discriminant Analysis, Logistic Analysis, etc.) and non-statistical techniques (Decision Trees, k-Nearest Neighbor, Cluster Analysis, Neural Networks, etc.). Each technique utilizes different assumptions and may or may not achieve similar results based upon the context of the data. Three of such models developed are regression methods, decision trees and artificial neural networks. Regression methods were termed as the study of dependence (97). This means it measures or calculates the relationship between dependent variable and one or more independent variables. Regression models are central part of many research projects. It has been used to predict the survival of critical conditioned patients who are generally admitted to intensive care unit as a function of physiological variables (98). Basically, regression models are classified into two main categories i.e. linear models and logistic regression models. The logistic regression model is quite often employed technique in data analysis. It is considered as a well-known classification modeling

that allows probabilistic decisions and shows promising results on several problems. Like all others regression models, which are used for description, control and prediction, logistic model (also called as logit model) produce similar results with a best fitting which is considered as a clinically interpretable model.

Survival analysis can be considered a classification problem in which the application of machine-learning methods is appropriate. By establishing meaningful intervals of time according to a particular situation, survival analysis can easily be seen as a classification problem. Survival analysis methods deals with waiting time, i.e. time till occurrence of an event. Commonly used method to classify this sort of data is logistic regression. Sometimes, the underlying assumptions of the model are not true. In model building, choosing an appropriate model depends on complexity and the characteristics of the data that affect the appropriateness of the model. Two such strategies, which are used nowadays frequently, are artificial neural network (ANN) and decision trees (DT), which needs a minimal assumption. This study aimed to compare predictions of the ANN, DT and logistic models by breast cancer survival.

6.2 Questions of Interest

Q1: What are the significant attributable variables which play an important role in classifying breast cancer survival?

Q2: What are the different models using different classification methods will be able to give improved prediction of survival in breast cancer women?

Q3: Which of the following techniques will produce the model with the highest precision in classifying the breast cancer survival data: logistic regression, decision trees, or neural networks?

Q4: How does ANN model and decision tree model perform compared with logistic regression model in the analyses breast cancer survival using different input variables for the same individuals?

Q5: Are there any benefits of Artificial Neural Network analyses (ANN) and decision tree models compared with logistic regression analyses?

Q6: Using the identified model, what is the probability of survived subject is correctly classified as survived and not survived woman as not survived?

Q7: Will the model selection vary based on the selected evaluation method?

6.3 Logistic Regression

The linear logistic regression assumes that natural logarithm of odds is in linear relationship with corresponding independent covariates. The linear logistic function is characterized by three main components. They are random experiment (identifies the PDF of response variable), a systematic component (linear relationship of explanatory variables which are used as predictors), link function (describes relationship between the first and second components). The logistic regression is distinguished from linear model based on its binary outcome. Logistic model is a type of predictive model which relates two categories of variables like dependent variables (dichotomous or binary outcome either 0 or 1) and independent variables (predictor or explanatory variables). In the binary response model, an individual takes one of the two possible outcomes. Some of the expected binary outcomes are active-inactive, healthy-unhealthy, normal-abnormal etc. For example the probability of officer promotion would relate to his characteristics like annual performance and CEP. This model estimates or predicts by fitting the occurrence of events into logistic curve. A broad choice of aspects using various

links functions that describe the relationship between the probability distribution of response variables and the linearity of explanatory variables are listed below.

1. The logistic function: $g_1(\pi) = log\{\pi/(1-\pi)\}$
2. The inverse normal function: $g_2(\pi) = \Phi - 1(\pi)$
3. The complementary log – log function: $g_3(\pi) = -log\{-log(1-\pi)\}$
4. The log – log function: $g_4(\pi) = -log\{-log(\pi)\}$

Apart from this logistic function also possess one important characteristic feature is its overall transformations in that it is eminently suited for analysis of data collected. Logistic regression architecture is given in Figure 6.1. For example, one can try to predict whether a subject will suffer from heart attack at a specified time based on certain characteristics like person age, sex, habitats etc. Logistic regression is extensively used in medical diagnosis like brain injury, different types of cancer prediction like breast, cervical, prostate etc. More details can be found in text book *Applied Logistic Regression* of Hosmer and Lemeshow (99). Example of logistic curve is shown in Figure 6.2.

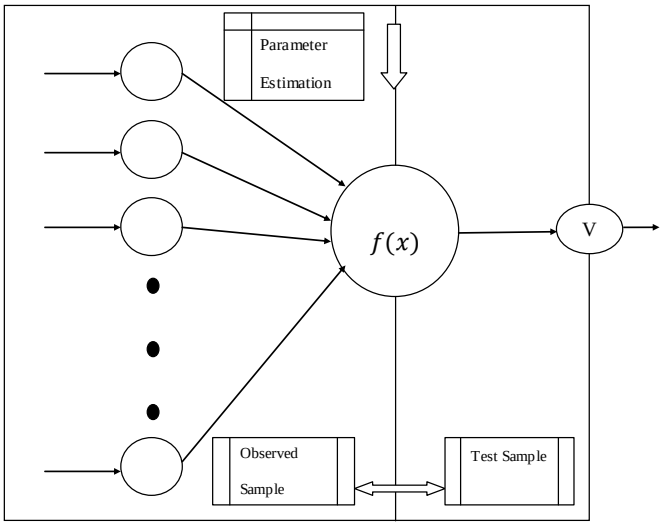


Figure 6.1 Architecture of Logistic regression

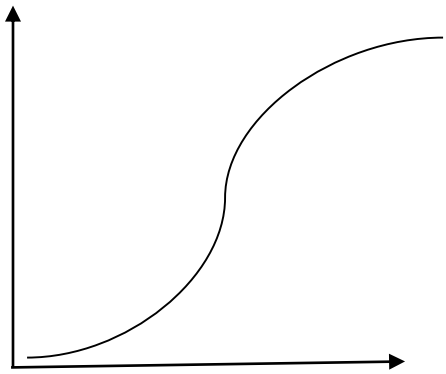


Figure 6.2 Logistic Curve

In case of polytomous response model, the response of a specific item or individual is restricted to only a fixed set of possible values. The binary response model falls under the category of polytomous response model as a special case. The logit models utilize the ordering of response variables by nature. One such example is usage of rating scales in testing of food and wine tasting.

McFadden (100) was the first person who linked the multinomial logit function to theory of mathematical psychology and received Nobel Prize in 2000. And many more articles in the 21st century have made their own and unique way of importance to logistic regression. At present wide range of applications using logistic function are being explored in various fields like medicine, biological sciences, sociology, psychology, business, management etc.

In our present work, the outcome variable, survival prediction with breast cancer or otherwise is predicted from the knowledge of the patient's age, tumor size, stage of cancer, treatment, administered and duration.

6.4 Timeline of Logistic Function

6.4.1: 19th Century

Alphonse Quetelet (1795-1874), Belgian astronomer turned statistician was first person who extrapolated the exponential growth of human population. Pierre-Francois Verhulst (1804-1847) derived the expression and named the expression as 'Logit function' (102). He included the expression, functions, properties and applications in three papers published at *Proceedings of the Belgian Royal Academy* (101).

6.4.2: 20th Century

1920-1930

Until 1920 there are no specific articles or reviews that discuss about logistic functions. Raymond Pearl and Lowell J Reed (1920) were the persons who discovered the logistic function for the study of population growth of United States of America. The curve gave a good fit for population during the period of 1790 to 1910. They do not have the knowledge of Verhulst works on Logit function. Berkson and Reed (1929) published papers on the application of logit function (103) to autocatalytic reactions in *Proceedings of the National Science Academy of Sciences*. Yule (1925) was the first person who provoked the name of logit function and appreciated the works of Verhulst in his papers in *Yule's Presidential Address* of the *Royal Statistical Society* (104).

1930-1940

Gaddum and Bliss (1933-1934) introduced the probit model also called as "Probability Unit". But the authors gave more importance to logarithmic transformations rather than common normal distributions in bioassay for the study of stimulus and its responses.

1940-1950

Berkson (1944) was the first person who substituted ‘Probit’ with ‘Logit’ by conducting many experiments on the method of maximum likelihood estimation and its advocacy in minimum chi-square estimation which were not approved at that time (105).Wilson(1943)was probably the first person to publish as application of the logistic function in bioassay in Wilson and Worcester.

1960-1970

Cox (1960-1970) gave equal importance to logit functions compared with probit functions in his articles published in JSTOR electronic repertory, which is one among the 12 major statistical journals in the English language. He covered the importance of multinomial generalization of logit function (37).

1970-1980

Mckelvey and Zavoina (1975) formulated the latent regression model for an ordered probit model of the voting behavior of United States congressmen (106). In 1977 BDR (Biomedical Data Processing) which is a computer package offered the facility of maximum likelihood estimation of logit and probit functions.

6.4.3: Recent Trends

Ever since the demand for logistic regression has increased tremendously, many articles in name and application of function evolved in many international journals. Few of much cited works are listed below for reference.

1991: The Importance of Assessing the fit of Logistic Regression Models (106).

1993: Nontraditional Regression Analysis (107)

1995: Regression Shrinkage and Selection via the Lasso (108)

1997: A Comparison of Goodness-of-Fit tests for the Logistic Regression Models (109).

1999: Additive Logistic Regression: A Statistical View of Boosting (110).

6.4.4 Underlying assumptions

There were many numbers of assumptions made to the logistic regression compared to ordinary regression methods.

8. The data collected is assumed to be completely randomized during the assignment of treatments to experimental subjects.
9. Multinomial logistic regression does not consider the sample size estimations and identifications of outliers.
10. The attracting aspect of multinomial logistic regression analysis is, it does not assume normality, linearity and homoscedasticity. In order to meet the requirements multinomial logistic regression is subjected to discriminant analysis because this analysis does not have any presumed assumptions.
11. The assumption of independent variables by logit function can be tested by McFadden-Hausman test (111).
12. Furthermore, Multinomial logistic regression assumes non-perfect separation which means if the outcomes of variables can be separated by predictor variables then unrealistic coefficients appear which influence the size.

6.4.5 Fitting the Logistic Regression Model and Significance Tests

Consider a sample size of *n* with observations $x_1, x_2, x_3, \ldots . x_n$which denote the predictor variables that produce the binary output either Y=0 (absence) or Y=1(presence) of the disease. 'Y' represents the dichotomous outcome variable corresponding to the x_i value of the i^{th}variable.

Assuming each of these variables is at least scaled interval, the conditional probability that is present denoted by $P(Y = 1|x) = \pi(x)$ where π denotes the probability of disease is present. The probability of outcome is related to the potential predictor variables by the equation of the form

$$logit\,[\pi(x)] = log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k$$

Where$(\alpha, \beta_1, \beta_2 \ldots \beta_n)$ are termed as the regression coefficients of the predictor variables $(x1, x2, x3 \ldots . xn)$. The coefficients of regression are extracted from the availability of data. The regression coefficients measure the percentage of contribution of predictor variables towards the outcome. This prediction is generally followed by the odds ratio of independent variable. The odds ratio is estimated by taking the exponential ratio of the coefficient (say:$exp\ (\beta 1)$). For example the odds ratio for breast cancer can be estimated by taking into consideration the age as independent variable along with exponential function of regression coefficient. This estimation represents the likelihood of occurrence of breast cancer based on age. The use of probability values determines the importance of variables in terms of statistical significance in producing outcomes. Increasing the sample size, predictors with small effects on the outcomes become statistically significant. Hence, the selection of significant variables is important in such a prediction. This selection is usually compelled either by forward or backward selection or step-wise selection depending upon the size of the sample. Sometimes clinically important variables may show statistically insignificant prediction of outcomes due to influence of strong predictors. In such case the criterion level of significance can be increased to avoid conflicts.

6.4.6 Survival prediction using Logistic, ANN and Decision tree modeling

In this chapter, using the same input and output variables we established four models using both logistic regression, ANNs, decision trees and compared their performances.

Event history models and logistic regression models are the two commonly used analyses of survival, where the former models use target survival as a continuous variable of survival time, while the latter models use a fixed survival length. The target is thus a dichotomous variable, survived or not. In this chapter, using logistic regression model as a classifier we predict the survival of breast cancer women.

The main idea of this chapter is to design four models with significant attributable variables to predict the survival of a breast cancer woman. The significant independent variables used in this modeling are selected by logistic regression analysis. As discussed earlier, logistic regression is a statistical technique used to examine the relationship between a dependent variable (survival or otherwise) and a one or more independent variables (numerical or categorical). Initially, we have used all the independent variables including: tumor size, age, stage of cancer, treatment, duration, grade of tumor, race, marital status, and number of primary tumors. Based on the logistic regression results the independent variables grade of tumor, race, marital status, and number of primary tumors nor their interaction terms were not statistically significant in providing the best prediction of survival of breast cancer women. Leaving these insignificant variables out of the modeling we designed four models inputting one variable at a time. The output vector in these models contains two variables for each case: predicted survival either 0 (not survived/dead) or 1 (survived/alive). A number between 0 and 1 gives an estimate of the accuracy of the predicted value.

The first model, named model-1 is using two variables including: age and tumor size only. The second model, named model-2 is using three variables including: age, tumor size, and stage of cancer. The third model, namely model-3 includes treatment along with the three variables chosen in model-2. The last model, model-4 has the variables including: age, tumor size, and stage of cancer, treatment and duration. In all these models our output is to predict the survival or otherwise of a breast cancer women. We interpret this overall accuracy, as a measure of the reliability of a given estimate.

Table 6.1 summarizes the specificity, sensitivity and overall accuracy results of the four logistic regression models. Table 6.2 has the ROC area values for the four logistic models. The results showed that the overall accuracy jumps from 70.42% for model-3 to 80% for model-4. This is not a surprising result. As anticipated, duration of stay for a woman with breast cancer, during the study period has a lot of importance for predicting accurate survival. The logistic regression model-1 yielded a ROC area of 68.8%, and sensitivity to survival of 95% gave a specificity of only 25%, model-2 with a ROC area of 71% and sensitivity to survival of 95% has a specificity of 30%. For the remaining two models the ROC area is 71.8% and 85.5% respectively and the sensitivity to survival of 95% has a specificity of 29% and 61% respectively. The results of model-4 logistic regression providing with overall accuracy of 80% along with 81.54% specificity, 76.82% sensitivity and 61% specificity at 95% sensitivity is often desirable. The sensitivity and specificity of all the four models with their respective confidence intervals are given in Table 6.1. For computing confidence intervals for sensitivity and specificity see Altman et al. The ROC graphs of the four logistic models are given in Figure 6.3.

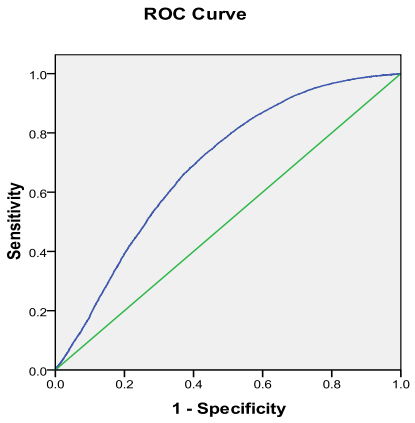
Table 6.1 Sensitivity, specificity and overall results of Logistic regression models

Logistic Regression Model	Sensitivity (%)		Specificity (%)		Accuracy (%)
	value	95% C.I	value	95% C.I	
LR 1	68.01	(67.04, 68.96)	69.45	(68.99, 69.91)	69.2
LR 2	67.31	(66.39, 68.21)	70.47	(70.0, 70.93)	69.78
LR 3	67.69	(66.80, 68.55)	71.26	(70.78, 71.72)	70.42
LR 4	76.82	(76.14, 77.47)	81.54	(81.11, 81.97)	79.98

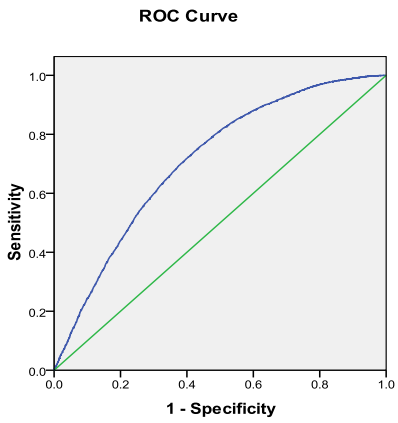
Table 6.2 LR models ROC area values

LR Models	ROC	At 95% sensitivity
		Specificity
LR-1	68.8%	25%
LR-2	71.0%	30%
LR-3	71.8%	29%
LR-4	85.5%	61%

LR-1



LR-2



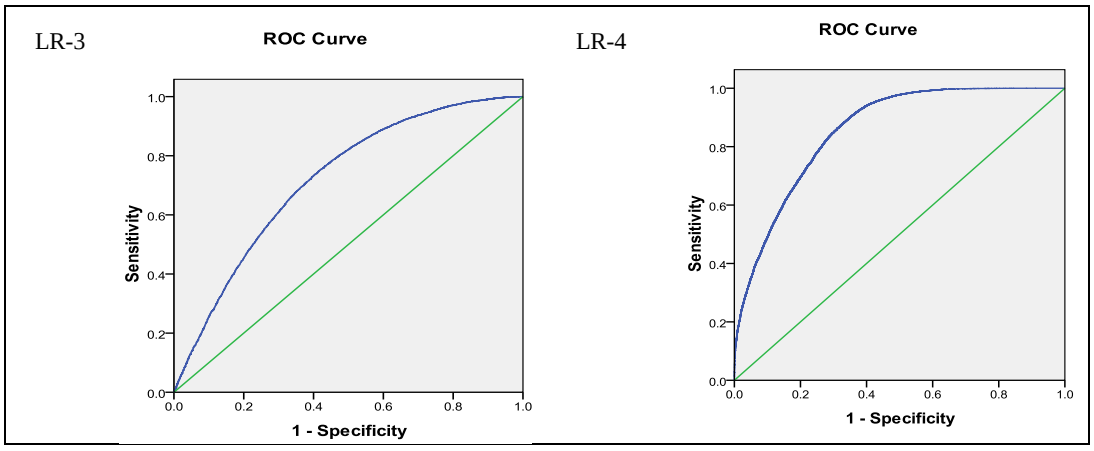


Figure 6.3 ROC graphs for four LR models

6.5 ANN Perceptron Classification

Major amount of research works during 1960's were carried under the name of "Perceptron". Frank Rosenblatt (1958) was the person who coined the term "Perceptron" in his psychological magazine (112). The word perceptron is derived from English word *"Perception"* which means ability of an individual to understand. He has written in his book named *"Principles of Neurodynamics"* on how to train these kinds of neurons to enable them perform pattern recognition tasks. He further provided information on how perceptron provide solution to particular problem in finite number of steps. The perceptron turns out to be McCulloch-Pitts model which mean a neuron with weighted outputs and with additional pre-processing.

6.5.1 Definition of Perceptron

A perceptron can be termed as a classification of different sets of data probably unseen data sets into learned ones. The structure of perceptron possesses a number of inputs, a bias and an output. A simple schematic diagram of perceptron is shown in Figure 6.4.

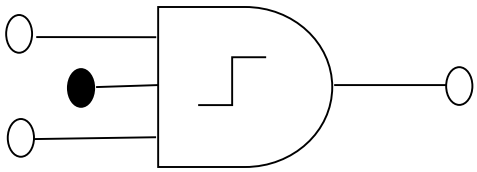


Figure 6.4 A simple perceptron

Another definition of perceptron can be considered as *"An arrangement of one input layer of McCulloh-Pits neurons feeding forward to one output layer of McCulloch-Pitts neurons is known as perceptron"*.

6.5.2 Multilayer Perceptron

The concept of multilayer perceptron is built using number of single layer neurons. Each of the perceptron layers is used to solve nonlinearly separable problems by breaking them into small linearly separable sections of inputs provided. The outputs of each individual perceptron is extracted and combined with another series of perceptrons to obtain final output. In most cases the hard-limiting function (step function) is used for producing outputs. This step function prevents the information of the inputs to overflow into the inner neurons. To solve this problem step function is replaced with a sigmoid function. In a multilayer perceptron, the neurons are arranged in order of the input layer, one or more hidden layers and an output layer as shown in Figure 6.5. The architecture (113) is designed to possess better properties like no direct connection between input and output layers, full connection between layers, number of outputs need not be equal to number of inputs, there is no limit for number of hidden layers i.e. they can be more or less than input and output units.

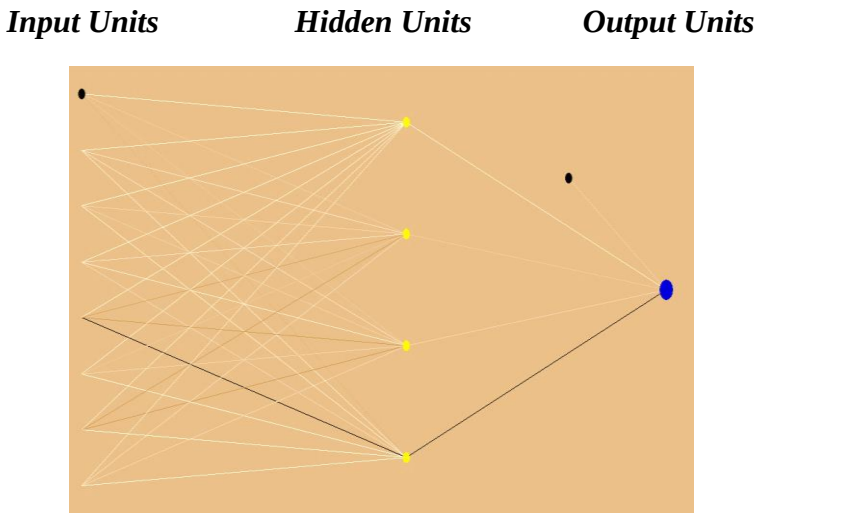


Figure 6.5 A simple feed forward perceptron model

ADALINE is an acronym for ADAtive LINear Element. Bernard Widrow and Marcian Hoff (1960) developed and presented this as single staged network. It is also called as the delta rule; the least-mean-squares rule; the Widrow Hoff rule. The binary values for input and output were assumed to be -1 and +1 respectively. Adaline possess similar architecture to perceptron, but the difference lies in type of learning rule used and thresholding step. These enable the user to solve the linearly inseparable problems which is impossible with single layer perceptron. The Widrow-Hoff learning (114) is applicable for trained supervisor, it is independent of the activation functions of neurons used. The LMS algorithm was proposed for Adaline. It is evident from above that training of perceptron requires modification of weights. The delta rule states that weights need to be adjusted corresponding to difference between desired and actual output.

6.5.3 Introduction to Back Propagation

The most widely used search technique for training artificial neural networks is back propagation. This can also be termed as “Feed-Forward back-propagation network”. This is a

user friendly model which can be understandable and implemented as software simulation. The development of the back propagation training algorithm was credited to Werbos (1993), Parker (1985) and LeCun (1986). It is considered as the most widely used learning which is easy to implement and train. Rumelhart, Hinton and Williams have made important contributions towards the development and analysis of back propagation (115). They have concentrated on the improvement of the original back propagation algorithm. The attempts include working on different strategies like scaling differentiation, error metric modification, transfer modification, architectural restructuring, and constraining the solution set of the problem.

The Back propagation is a local search technique which is still a popular and successful tool. It requires training for conditioning the network before used for processing other data. Networks possess one or more hidden layers depending upon the training introduced. Supervised training provides preliminary adjustments to the weights associated to organize the patterns categorically. Even though BP is most popular optimizing method to train networks it has certain limitations like inconsistency and unpredictable performances. The gradient nature of BP could be eliminated by using global search techniques which do not depend on their derivatives. There are some cases where large networks can take long time to be trained and may not converge to solution significantly. The building of neural network ideal to brain is impossible. However we can build some simpler artificial neural networks with a suitable transfer function to work almost similar to a biological neuron. The functions of neural network built works similar in meaning to the human brain.

The Feed Forward, Back-Propagation architecture (116) was developed in the early 1970's by various independent sources (Werbor; Parker; Rumelhart, Hinton and Williams). In Feed-Forward propagation, neurons in present layer receive signals from preceding layers which

is multiplied by corresponding weights separately. Inputs from one or more previous neurons are individually weighted, then summed. The entire uniqueness of the network exists in the values of the weights between neurons. For this type of network in order to adjust weights the most common learning algorithm is called back propagation (BP).The use of term "Back Propagation" appears to be evolved after 1986 when researchers have presented their research of results on *Parallel Distributed Processing* (PDP) models. This synergistically developed back-propagation architecture which is most effective and easy to learn model for multilayer networks. Some work has been done which indicates that a maximum of five layers, one input layer, three hidden layers and an output layer are required to solve problems of complexity.

6.5.3.1 Training with back propagation

The problems are classified into training, testing and validation, files in the description of data sets. A BP network will search for a solution using the training data, if the error decreases during the testing & validation step, the training will discontinue. The researchers believe this step is necessary to not over fit a particular function being estimated. The problem of the algorithm begins with convergence. It may either converge to local or global solution. If a correct objective function is chosen and a global solution is obtained, then there will no such problem. Since, BP converges locally this type of NN training seems to be necessary. Learning rate (training parameter that controls the size of weight and bias changes during learning) and momentum coefficient (used to prevent the system from converging to a local minimum or saddle point) are the key factors that will help a network to train. Too low a learning rate makes the network learn very slowly. Too high a learning rate makes the weights and objective function diverge, so there is no learning at all. In training our networks we set the learning rate as 0.15 and the momentum as 0.8.

6.5.3.2 Back-Propagation Algorithm

A gradient search technique (117) like BP can provide the user with well recognized problem such as escaping local optima. The weights which are initialized randomly during training and starting point is located in local valley with high probability. Numerous solutions have been proposed to problems like differential scaling, the transfer function etc. assuming many different random starting points. A user must be able to choose different parameters to apply in neural networks software packages. The parameters include step size, momentum, learning rule, normalization technique, random seed etc. to find best combination to solve a particular problem. For the training of multilayer feed-forward ANNs, Error-Back propagation algorithm plays an important role. Generally the input layer is considered as a just distributor of signals from the external world and not taken into consideration as a layer.

The back propagation training consists of two methods of computation:

1. A forward pass
2. A backward pass

In forward pass an input pattern vector to the units in the input layer basically leads to the sensory nodes of the network. The signals from the input layer then propagate to series of layers finally producing the output. This process continuous until the signals reach output layer where actual response of the network to the input vector is obtained. In the backward pass, the synaptic weights are adjusted according to the signal which propagated backwards to the direction of the synaptic connections.

6.5.3.3 Implementing Back Propagation

The back propagation algorithm can be implemented in two different modes:

1. On-line mode

2. Batch mode

In the on-line mode the error function is calculated after the presentation of the input pattern and the error signal is propagated back through the network modifying the weights before the presentations of the next pattern. The error function is generally the Mean Square Error of the difference between the desired and the actual responses of the network. All such presentations of such patterns is usually called as an epoch or one iteration. In batch mode the weights are modified only when the input pattern have been presented. Then the error function is calculated as the sum of the individual MSE for each of the input pattern and weights are modified accordingly before the next iteration.

6.5.4 Error functions

If a pattern is submitted and its classification or association is determined to be erroneous, the synaptic weights as well as the thresholds are adjusted so that the current least mean square classification error is reduced. The input - output mapping, comparison of target and actual values, and adjustment, if needed, continue until all mapping examples from the training set are learned within an acceptable overall error. Usually, mapping error is cumulative and computed over the full training set. Error is the measure of the discrepancy between the neural network output and the target. The most popular error functions are sum of squares (SSE) and cross entropy (CE) among others.

6.5.5 Advantages of Multilayer Perceptrons

The general characteristics of multilayer perceptrons are generalization and fault tolerance.

Generalization: Neural networks are capable of classifying unknown patterns with the support of known patterns that have some different level of features. This means incomplete inputs will be classified because of their similarity with complete inputs.

Fault Tolerance: Neural networks are highly fault tolerant. This characteristic feature can also be termed as "graceful degradation" (118). Hence the neural networks keep on working even if some interconnections between some neurons fail.

6.5.6 Limitations of Multilayer Perceptrons

There are limitations to the feed forward, back propagation architecture. Back-propagation requires a lot of supervised training, with lots of input-output examples. Sometimes, the learning can get stuck in local minima, limiting the best solution. This occurs when the network systems finds an error that is lower than the surrounding possibilities but does not finally gets to the smallest possible error. In typical feed forward, back-propagation applications, the desired output may not be known precisely. In such case the back propagation learning cannot be used directly. Examples like include speech synthesis from the text robot arms, evaluation of bank loans, image processing etc.

6.5.7 ANN Modeling

Neural networks are undoubtedly powerful nonlinear function estimators. As mentioned earlier there are several types of ANN architectures. They usually perform prediction tasks at least as well as other techniques, if not significantly better. Additionally, building an ANN requires minimum domain knowledge in the areas of mathematics and statistics, than does for building a logistic regression model. The ANN type used in this study is called a multilayer perceptron (MLP) or multilayer feed forward network, which propagates input signals forwards and error signals backwards. During the process, the weights are adjusted so that the output

grows more accurate. This process is prone to over fitting problems. In order to avoid over fitting, a common technique is to train the network with some portion of the data values, and then evaluate its performance by testing the trained network with the remaining data values. In our ANN modeling we used 70% data for training and remaining 30% data for testing.

The four ANN models consisted of an input layer, a hidden layer and an output layer. Table 6.3 summarizes the specificity, sensitivity and overall accuracy results of the ANN models when training. Table 6.4 summarizes the specificity, sensitivity and overall accuracy results of the ANN models when testing the trained model. Table 6.5 has the ROC area values for the four ANN models. Since training is the key factor for an ANN model, here we will be discussing about training results of ANN models. Even in this case, the results showed that the overall accuracy jumps from 71.12% for model-3 to 82.80% for model-4 for the same reason as mentioned earlier. The ANN model-1 yielded a ROC area of 72.1%, and sensitivity to survival of 95% gave a specificity of only 31%, model-2 with a ROC area of 73.1% and sensitivity to survival of 95% has a specificity of 32%. For the remaining two models the ROC area is 73.8% and 87.4% respectively and the sensitivity to survival of 95% has a specificity of 39% and 66% respectively. Comparing these results with logistic models, at a 95% sensitivity, ANN has a better specificity for all the four models. Table 6.5 gives the details about architecture and ROC area of ANN models and their respective ROC graphs of the four ANN models are given in Figure 6.7. The output of ANN will be a composite function of the form

$$y_i = f\left\{\sum tanh\left(\sum(\bullet)\right)\right\};\ i = 0,1;$$

$$f(\bullet)\ is\ a\ softmax\ function\ and\ tanh(\bullet) is\ a\ hyperbolic\ tangent\ function$$

Table 6.3 Sensitivity, specificity and overall results of ANN training

ANN Models	Sensitivity (%)		Specificity (%)		Accuracy (%)
	Value	95% C.I	Value	95% C.I	
ANN 1	66.78	(65.70, 67.83)	70.76	(70.19, 71.31)	69.85
ANN 2	67.25	(66.19, 68.27)	71.48	(70.91, 72.03)	70.46
ANN 3	68.23	(67.20, 69.23)	72.08	(71.51, 72.63)	71.12
ANN 4	88.95	(88.27, 89.59)	80.60	(80.09, 81.09)	82.80

Table 6.4 ANN models architecture and ROC values

ANN Models	Architecture	ROC	At 95% sensitivity
	I – H - O		Specificity
ANN-1	2 – 7 – 2	72.1%	30%
ANN-2	6 – 3 – 2	73.1%	32%
ANN-3	10 – 6 – 2	73.8%	39%
ANN-4	11 – 3 – 2	87.4%	66%

Table 6.5 Sensitivity, specificity and overall results of ANN testing

ANN Model	Sensitivity (%)		Specificity (%)		Accuracy (%)
	value	95% C.I	value	95% C.I	
ANN 1	66.18	(64.52, 67.80)	70.66	(69.79, 71.51)	69.63
ANN 2	68.66	(67.08, 70.20)	72.00	(71.14, 72.84)	71.20
ANN 3	66.87	(65.26, 68.43)	71.89	(58.27, 59.97)	70.67
ANN 4	89.36	(88.32, 90.32)	80.97	(80.20, 81.72)	83.20

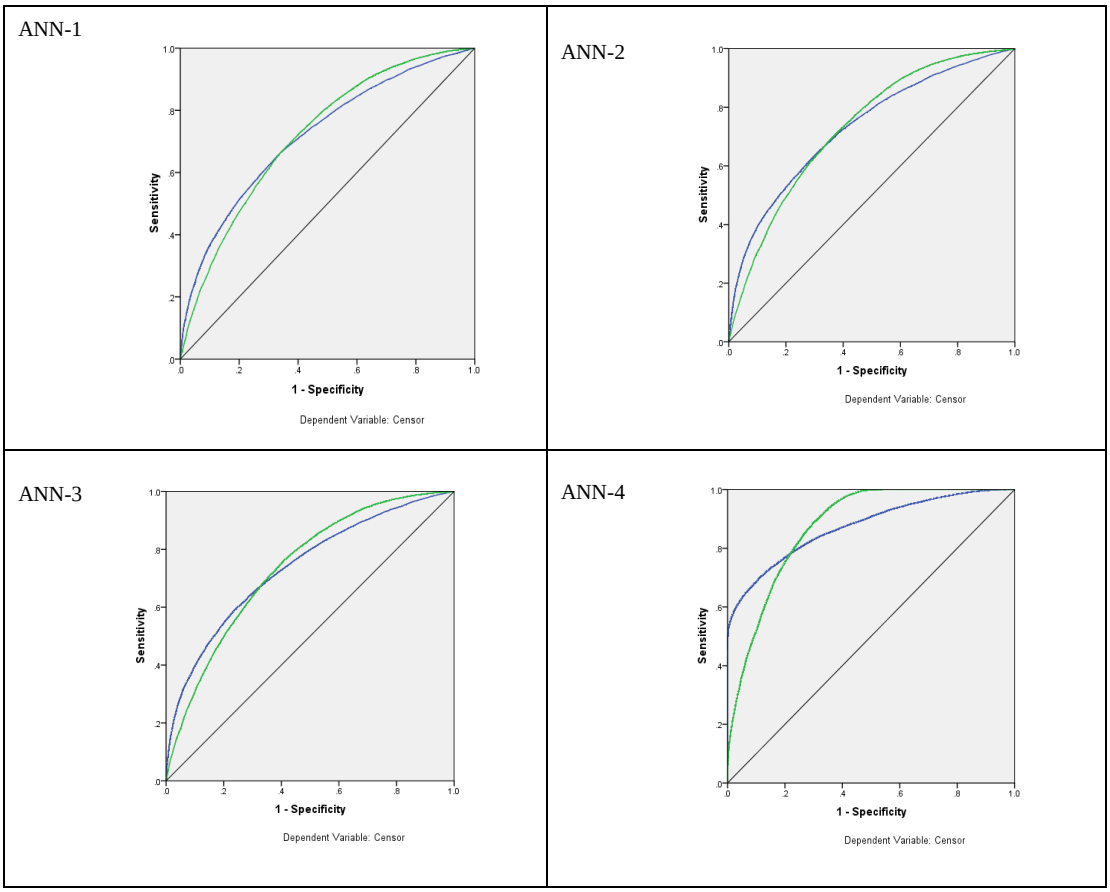


Figure 6.6 ROC graphs for four ANN models

6.6 Decision Tree Classification

Data mining tools are proved to be successful in field of medical diagnosis. The combination of both data mining tools along with decision trees is popular and effective classification approach which provides understandable and clear classifications rules that transfer knowledge to physicians and medical specialists. Data mining methods help to reduce the false positive and false negative decisions (129-131). This is one of the actively employed techniques that provide promising results in the breast cancer diagnosis.

A decision tree can be stated as the classification tool or classifier for determining appropriate action for the given situation. A simple decision tree consists of a root node (parental node), internal nodes or test nodes, and leaf nodes (terminal nodes or decision nodes). The final decisions for the target class are obtained on the leaf nodes from performing split test in the internal nodes. In complex cases, the leaf node possesses a probability vector for the target value of certain case (132). A simple decision tree classifying survival of breast cancer patients with treatment as an attributable variable is given below Figure 6.7.

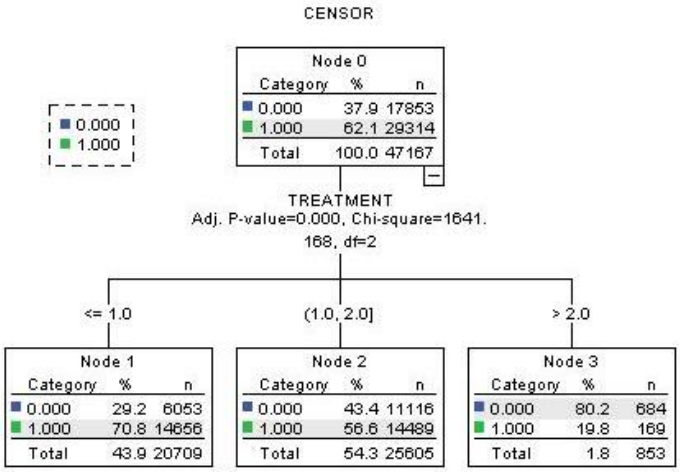


Figure 6.7 Simple Decision Tree example

Decision tree usually consists of nominal and/or continuous attributes. In case of nominal attributes, one outcome is assigned for the target value whereas for continuous attributes there will be threshold which has two outcomes, one for each classified interval based on the conditions imposed by the fixed threshold. A more comprehensible decision trees are typically less complex preferred by the decision makers. Each designated path of the decision tree from root to leaf can be transformed into a rule by computing tests along the path which assign class prediction to terminal node. These predictions are termed as the class values.

6.6.1 Framework of Decision Trees: Algorithm

Decision tree accuracy is affected by the total number of nodes, depth of the tree, total number of leaves and number of attributes used. The complexity is controlled explicitly by the stopping criteria used and pruning methods employed. The objective of the decision trees is to find the optimal decision tree by minimizing the general errors. In order to solve heuristic problems with large data sets decision tree inducers with growing and pruning are being actively employed. The algorithms employed follow the concept of "divide and rule" in evaluating for the final optimal decision tree. In the foregoing process, partition of the training sets is executed based in the values of the discrete attributes. The appropriate function is selected based on the splitting measures. After the selection, nodes are further divided into subsections to carry out similar splitting procedures or stopped when the criteria is satisfied (133, 134).

6.6.2 Splitting Techniques

Decision trees are most commonly univariate splitting i.e., they make splitting measures based on the single attribute at each internal node. But, the inducer searches for the best attribute at internal node upon splitting. Various criteria contain measures for the splitting procedures to be executed. The splitting procedures are employed in different ways based on the originating measure (includes information theory, dependence and distance) and based on the measure of structure (impurity based criteria, normalized impurity based criteria, and binary criteria) more of which can be found in data mining books (135). In case of univariate splitting, many researchers claim that the choice of splitting criteria does not make much difference on the performance of the tree.

Accordingly in the literature multivariate splits have been extensively employed in case complex decision making situations. The frame work for these splits is not well known as that of univariate splits. Several attributes are involved in the single node split test at each internal node. Generally, multivariate splits are based on the linear combination of the input variables. The problem of finding the optimal linear split is much more difficult than that of the univariate split. Methods used for finding optimal split include greedy search method (136), linear programming (137), linear discriminant analysis (138), and many others.

6.6.3 Stopping Criteria

All the decision trees require stopping criteria otherwise it would be an undesirable to grow a tree which occupies its own node. This would lead to expensive computation and difficulty in interpretation. Rules for stopping the growing phase are discussed below.

1. Number of cases in the node is less than the pre-indicated value.
2. The depth of the node should not exceed more that predefined or maximum value.
3. The number of cases in the terminal nodes is less than the minimum number of cases for parent nodes.
4. The best splitting should not exceed a certain threshold limit set.
5. Predictor values for all records are identical – no further rule for splitting is computed.

6.6.4 Pruning Methods

Early studies have proved stopping criteria degrade the performance of tree. This might create small and under fitted trees or over-fitted trees depending on situations. Hence, an alternative method for stopping growth is to allow the tree to grow and prune back to the optimum size using certain pruning methods. Pruning methods gained importance based on

trading accuracy for simplicity. It has improved the generalized performance of the decision tree especially in noisy circumstances (139). There are various techniques for pruning the trees include cost-complexion, reduced and minimum error, pessimistic, optimal etc.

6.6.5 Decision Tree Inducers

The approach of induction is to develop a decision tree from set of examples. Various techniques like ID3, C4.5, Classification and regression trees (CART or CRT), chi-squared automatic interaction detector (CHAID), Quick, unbiased, efficient, and statistical tree (QUEST) and many others are actively employed based on the attributes. For large data sets two methods developed have been popularly employed namely the Catlett method and SLIQ algorithm. Further advancements and extensions for decision trees like oblivious trees (140), fuzzy decision trees, and incremental induction (141) can be found in the literature. Here in this chapter we will construct decision trees based on CHAID and CRT methods and choose the best performing method.

6.6.6 Chi-squared Automatic Interaction Detector (CHAID)

CHAID is a type of decision tree technique, based upon adjusted significance testing (Bonferroni testing). It is one of the oldest tree classification methods originally proposed by Kass (1980; according to Ripley, 1996, the CHAID algorithm is a descendent of THAID developed by Morgan and Messenger). . CHAID algorithm only accepts nominal or ordinal categorical predictors. When predictors are continuous, they are transformed into ordinal predictors before using the following algorithm. After the merging of the continuous and categorical variables adjusted p-value is computed using Bonferroni adjustments (14). P-value decides further merging operation if needed or not.

6.6.7 Classification and Regression Trees (CART)

CART algorithm was introduced in Breiman in 1986. These trees are characterized by the construction of binary trees implies that each external nodes consists of exactly two outgoing edges. It generates a regression model when the target variable is continuous else a classification model in case of categorical variables. In case of regression models, the CART looks for the splits that minimize the prediction square error (8). The prediction is based on the mean value of the target attribute of the rows falling under the terminal leaf node. The present research studies have employed these two methods which appeared to give better results compared to other evaluation methods.

6.6.8 Advantages and Disadvantages

Decision trees were pointed as good classification tools in literature due to its self-explanatory nature and easy to understand and interpretation behavior. It takes into consideration both numerical and nominal input attributes. They have the capability to handle and deal with large datasets and datasets with large amount of errors. The predicted performance is proved to be much higher and better than traditional methods like neural networks, logistic methods etc.

Decision trees also possess certain disadvantages which include its sensitiveness to small changes in input data can alter the nature of trees. Most of the algorithms accept only discrete variables (like ID3 and C4.5). Decision trees perform well if few highly relevant attributes are present and less if more complexions interactions exists.

6.6.9 Modeling using Decision Trees

Table 6.6 summarizes the specificity, sensitivity and overall accuracy results of the four decision tree models using both CHAID and CRT based methods. Table 6.7 has the ROC area

values for these models. The results showed that the overall accuracy jumps from 71.10% for model-3 to 82.6% for model-4 in a CHAID decision tree. Similarly for a CRT based decision tree the accuracy jumps from 70.9% for model-3 to 83.2% for model-4. As noticed in both logistic and ANN models, duration under study for a woman with breast cancer, during the study period has a lot of importance for predicting accurate survival.

For CHAID based decision tree the ROC for model-1 covered an area of 72%, and sensitivity to survival of 95% gave a specificity of only 22%, model-2 with a ROC area of 73.2% and sensitivity to survival of 95% has a specificity of 25%. For the remaining two models the ROC area is 73.6% and 87.6% respectively and the sensitivity to survival of 95% has a specificity of 29% and 62% respectively. The results of model-4 decision tree with overall accuracy of 82.6% along with 80.14% specificity, 89.67% sensitivity and 62% specificity at 95% sensitivity are often desirable. The sensitivity and specificity of all the four models with their respective confidence intervals are given in Table 6.6. For computing confidence intervals for sensitivity and specificity see Altman et al. The ROC graphs of CHAID based decision tree models are given in Figure 6.9.

The results of CRT based decision tree models reported a ROC of 71.9% for model-1, and sensitivity to survival of 95% gave a specificity of only 24%, model-2 with a ROC area of 72.8% and sensitivity to survival of 95% has a specificity of 29%. For the remaining two models the ROC area is 72.7% and 87.4% respectively and the sensitivity to survival of 95% has a specificity of 28% and 62% respectively. The results of model-4 CRT decision tree with overall accuracy of 82.2% along with 79.86% specificity, 93.62% sensitivity and 62% specificity at 95% sensitivity are often desirable. The sensitivity and specificity of all the four models with their

respective confidence intervals are given in Table 6.6. For computing confidence intervals for sensitivity and specificity see Altman et al. The ROC graphs of CRT based decision tree models are given in Figure 6.8.

CHAID uses multi way splits by default (meaning that a given current node is split into more than two nodes), whereas CRT does binary splits (meaning each node is split into two sub-nodes only). This difference between CRT and the CHAID has even an effect on the tree structures. In case of CHAID, trees sometimes look more like bushes. CHAID has been especially popular in marketing and medical research, where segmentation or classification has many major applications. Few more differences are listed below:

- CHAID uses a p-value from a chi-square significance test to measure the desirability of a split, while CRT uses the reduction of an impurity measure.
- CHAID searches for multi-way splits, while CRT performs only binary splits.
- CHAID uses a forward stopping rule to grow a tree, while CRT deliberately over fits and uses validation data to prune back.
- CHAID tree output is simple, short and easy to interpret, while CRT has a larger tree structure.

Finally, one may prefer CHAID when the goal is to classify or understand the relationship between a response variable and a set of explanatory variables, whereas CRT is better suited for creating a regression model. In view of this, in this chapter we will choose CHAID over CRT for survival classification of breast cancer woman.

Table 6.6 Sensitivity, specificity and overall results of Decision trees

Training		**Sensitivity**		**Specificity**		**Accuracy (%)**
		Value	95% CI	Value	95% CI	
Model 1	CHAID	64.77	(63.73,65.8)	71.12	(70.60,71.73)	69.6
	CRT	66.83	(65.74,67.91)	70.75	(70.19,71.30)	69.9
Model 2	CHAID	69.33	(68.27,70.37)	71.21	(70.65,71.76)	70.8
	CRT	69.2	(68.14,70.24)	71.2	(70.63,71.74)	70.7
Model 3	CHAID	67.59	(66.58,68.59)	72.25	(71.69,72.81)	71.1
	CRT	66.22	(65.22,67.21)	72.64	(72.07,73.20)	70.9
Model 4	CHAID	89.67	(89.00,90.30)	80.14	(79.63,80.63)	82.6
	CRT	93.62	(93.05,94.14)	79.86	(79.35,80.35)	83.2
Testing		**Sensitivity**		**Specificity**		**Accuracy (%)**
		Value	95% CI	Value	95% CI	
Model 1	CHAID	63.66	(62.07,65.22)	71.22	(70.35,72.09)	69.3
	CRT	67.33	(65.66,68.97)	69.69	(68.83,70.55)	69.2
Model 2	CHAID	68.63	(66.97,70.24)	71.27	(70.40,72.12)	70.7
	CRT	67.24	(65.56,68.87)	70.82	(69.95,71.67)	70
Model 3	CHAID	66.32	(64.75,67.85)	72.19	(71.31,73.04)	70.7
	CRT	65.11	(63.55,66.65)	71.42	(70.54,72.28)	69.8
Model 4	CHAID	88.2	(87.10,89.21)	80.38	(79.61,81.14)	82.4
	CRT	94.02	(93.19,94.76)	79.93	(79.16,80.69)	83.2

Table 6.7 ROC of Decision tree using CHAID and CRT

	CHAID	CRT
Model-1	72.0%	71.9%
Model-2	73.2%	72.8%
Model-3	73.6%	72.7%
Model-4	87.6%	87.4%

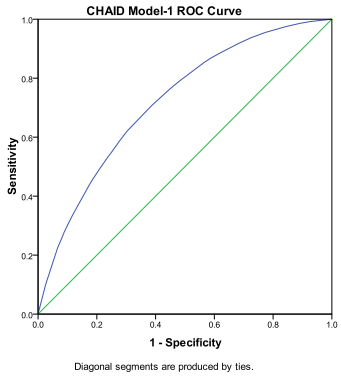
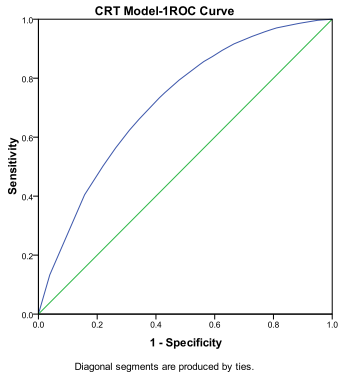
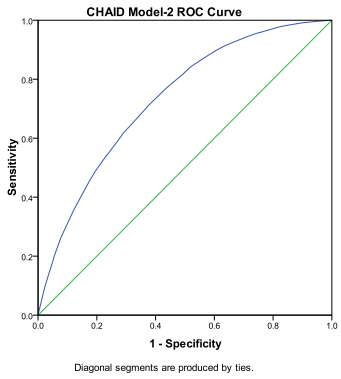
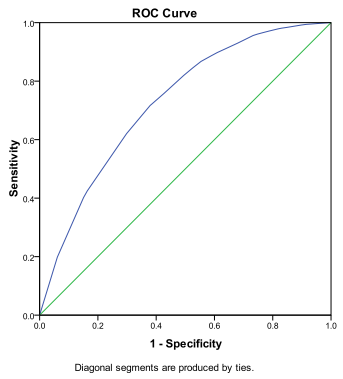
6.7 Performance Evaluation of models

In the context of predictive binary classification models, one of four outcomes is possible: (a) a true positive (TP) – i.e., a survived subject is classified as “survived”; (b) a false positive (FP) – i.e., a not survived subject is classified as “survived”; (c) a true negative (TN) – i.e., a not

survived subject are classified as “not survived”; (d) a false negative (FN) – i.e., a not survived subject is classified as “survived”.

The central concern of implementing different modeling applications in this chapter is to identify which of the proposed techniques are actually improving predictive accuracy. An improvement of even a fraction of a percent can translate into significant savings or increased revenue.

The performances of logistic, ANN and decision tree models in this chapter are evaluated based on the sensitivity, specificity, overall accuracy, and the area under curve values of each model. Sensitivity is the proportion of true positives that are correctly identified by the model. Specificity is the proportion of true negatives that are correctly identified by the model.

	CHAID	CRT
Model-**1**		
Model-**2**		

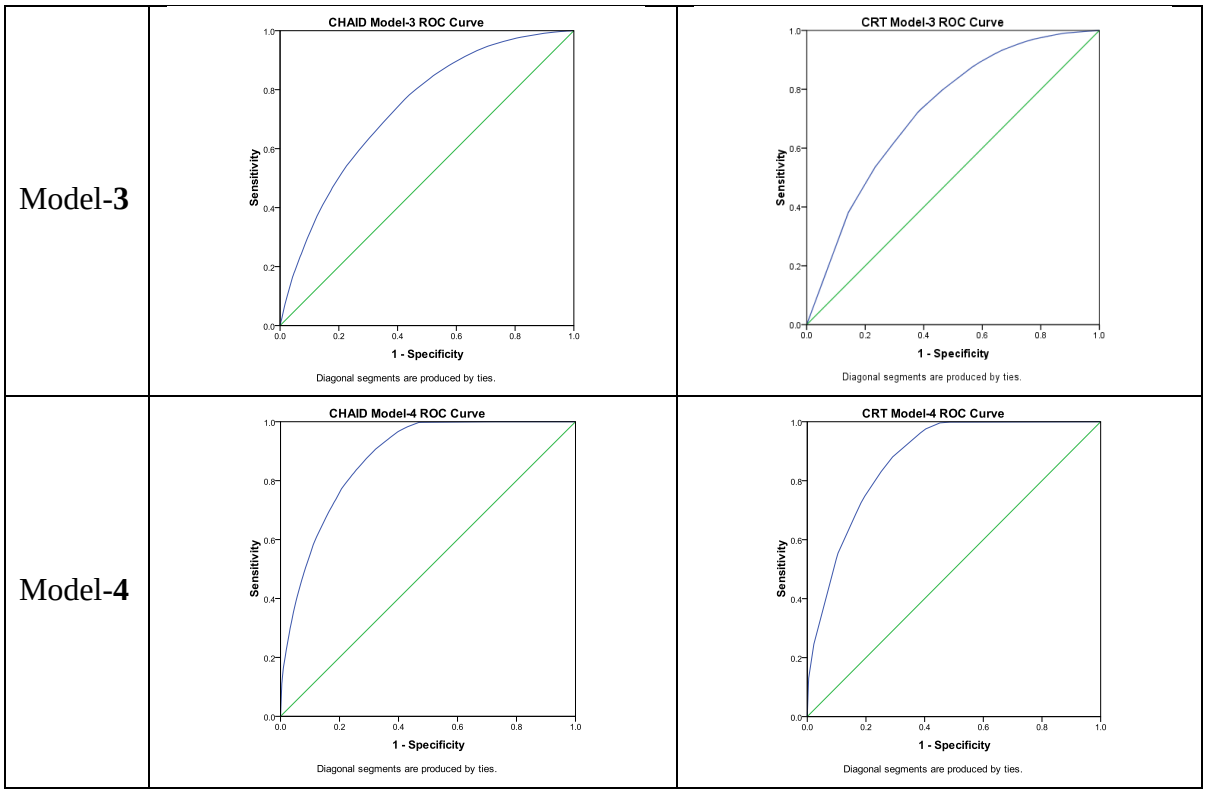


Figure 6.8 ROCs of Decision trees using CHAID and CRT

In other words, sensitivity refers to how good is the designed model is at correctly identifying women who are dead with breast cancer and specificity refers to how good the designed model is at correctly identifying women who have survived breast cancer (119). However, as a matter of fact, reporting a high sensitivity is not necessarily a good thing, but it's the specificity, which should not be worse, which in turn can conclude the designed model as useless (120). Also, we will compare the area under the ROC curve, which is a convenient way to compare different predictive binary classification models when the analyst or decision maker has no information regarding the costs or severity of classification errors. According to Thomas (2000), this measurement is equivalent to the Gini index and the Mann-Whitney-Wilcoxon test statistic for comparing two distributions (Hanley and McNeil) and is referred in the literature in many ways, including AUC or AUCROC values.

Many research studies have exhibited the importance of ANNs, decision trees, logistic regression as predictor and classification tools in field of medical diagnosis. The works are extended in the risk prediction in a variety of cancers like breast (121), prostate (122), liver (123), ovarian (124), cervical (125), bladder (126), and skin cancer (127).

We will compare the results of four logistic models with the results of four ANN models and decision tree models. The analytical description of designed neural network or the internal working of the ANN models will not be our point of concentration however we will treat them as black box which intakes input data and gives us the output.

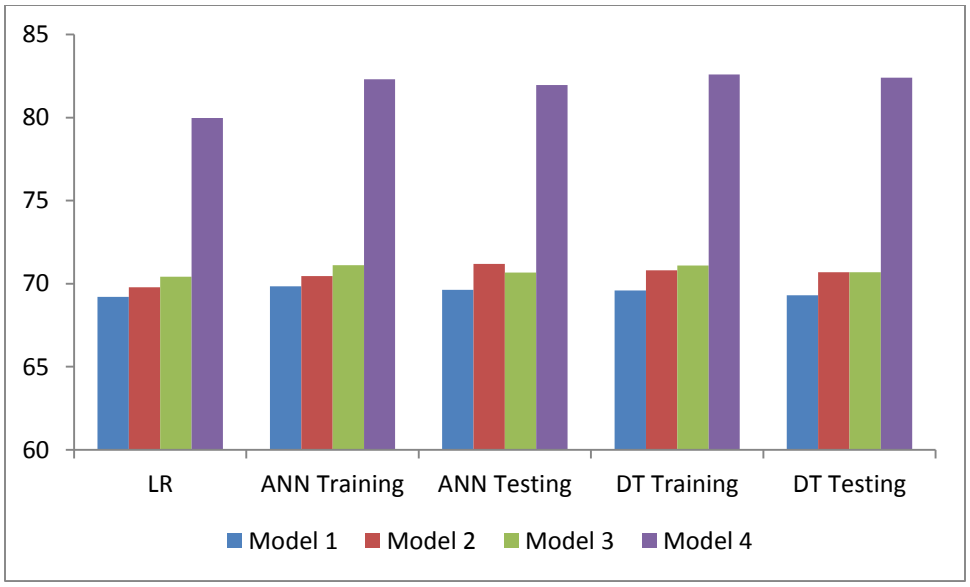
Table 6.8 has the performance evaluation of logistic, ANN and decision tree techniques. The overall accuracy for correct classification of survival of breast cancer women is almost the same in ANN and decision tree techniques compared to logistic. However the specificity of the model performance for logistic is slightly more than the ANN and decision trees. The ranking of these methods based on their classification performances are also tabulated in Table 6.8. The area under the curve ROC values of decision tree methods is slightly more compared to ANN and logistic regression methods. Table 6.9 has the details of comparing ROCs of the three different methods employed in this chapter with their ranking based on high ROC values.

Table 6.8 Performance Comparison of Logistic, ANN and Decision tree

	Overall Accuracy					**Specificity**				
Model	**LR**	**ANN**		**CHAID**		**LR**	**ANN**		**CHAID**	
		Train	Test	Train	Test		Train	Test	Train	Test
1	69.2	69.85	69.63	69.6	69.3	69.45	70.76	70.66	71.17	71.22
2	69.78	70.46	71.2	70.8	70.7	70.47	71.48	72	68.63	71.27
3	70.42	71.12	70.67	71.1	70.7	71.26	72.08	71.89	72.25	72.19
4	79.98	82.31	81.95	82.6	82.4	81.54	79.76	79.22	80.14	80.4
Rank	**III**	**II**		**I**		**I**	**III**		**II**	

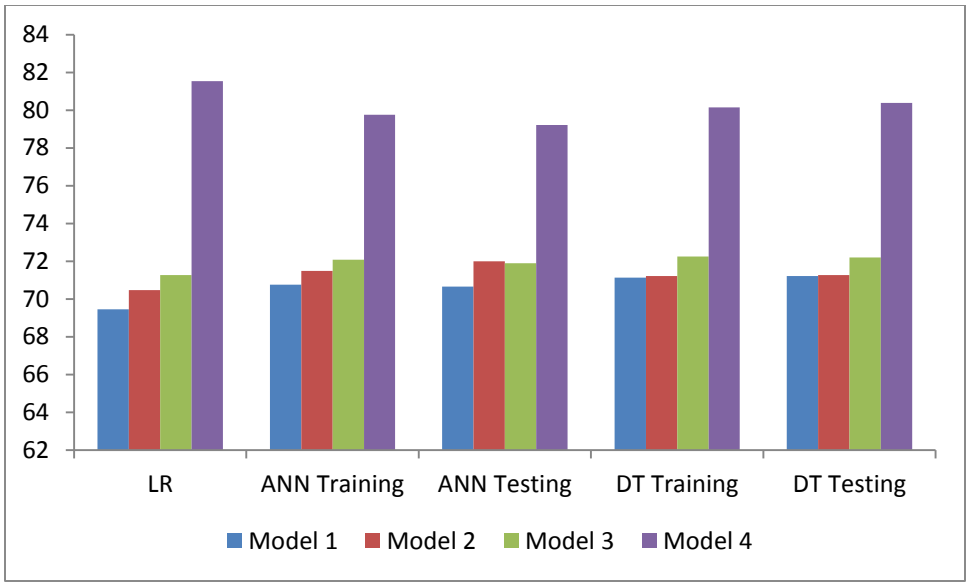
Table 6.9 ROCs of all methods

Models	LR	ANN	DT
Model-1	68.8%	72.1%	72.0%
Model-2	71.0%	73.1%	73.2%
Model-3	71.8%	73.8%	73.6%
Model-4	85.5%	87.4%	87.6%
Rank	**III**	**II**	**I**



LR-Logistic Regression; ANN-Artificial Neural Network; DT-Decision tree

Figure 6.9 Comparison of overall accuracy of LR and ANN models



LR-Logistic Regression; ANN-Artificial Neural Network; DT-Decision tree

Figure 6.10 Specificity comparison of LR and ANN models

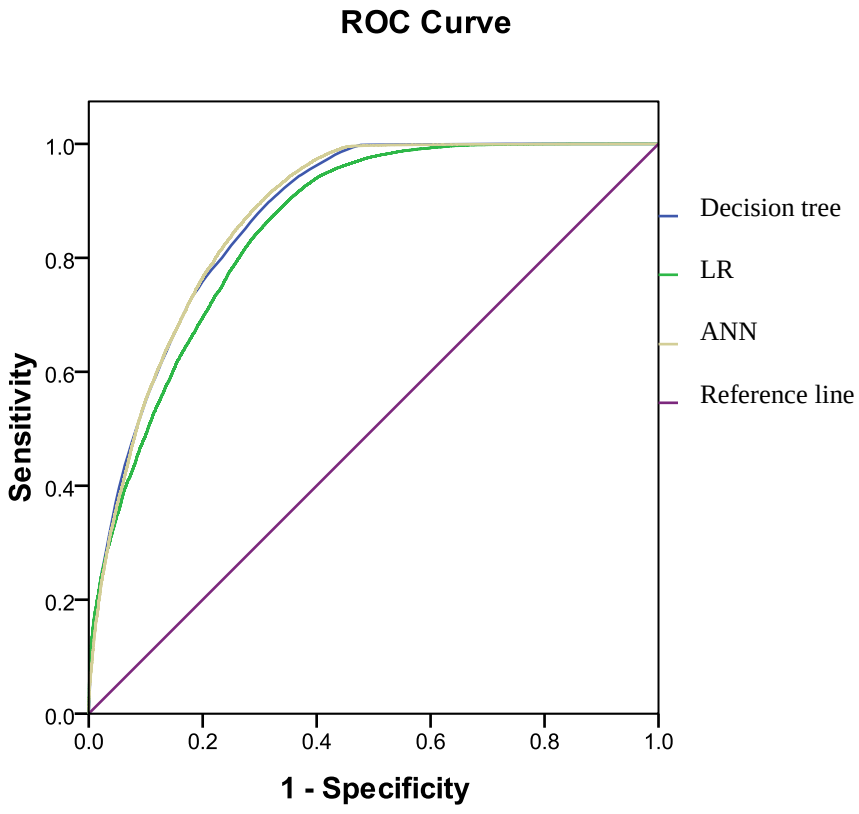


Figure 6.11 Comparison of ROCs graphically for the three methods

6.8 Conclusion and discussion

For maintaining consistency in comparing the models, we initially compared the accuracy of logistic distribution and accuracy in ANN and decision tree models in classification of survival of breast cancer data. Further we calculated the receiver operating characteristic (ROC) curves, compared them visually and calculate the area under the curve for comparison (121). The graph of comparison of three methods based on ROCs is given in Figure 6.11. Model-4 using the inputs including age, tumor size, stage of cancer, treatment and duration performed well by logistic, ANN and decision tree methods. The accuracy of classification for LR, ANN and DT models is recorded as 79.98%, 82.31% and 82.6%. We find no much difference in these values for ANN and DT methods. However, at 95% sensitivity ANN has reported a better specificity compared to logistic and DT models. Using ROC analysis as a measure of discriminating ability of logistic, ANN and decision tree models we have not found convincing proof that the use of ANN model or decision tree models in general would increase the quality of the statistical studies that use traditional tools such as logistic regression models.

In the present study, the effects of factors like age, tumor size, stage of cancer, treatment, and on the survival of a woman with breast cancer were designed. Four models for logistic and four models of ANN trained with gradient descent and four models of DT based on CHAID algorithm have been evaluated. The degree of generalization or the precision of predictive ability was measured for each logistic model, ANN model, decision tree model and their predictive abilities were in the order of model-4 > model-3 > model-2 > model-1.

As mentioned earlier, there is no significant difference in performance between LR, ANN model and decision tree models as measured by area under the ROC curve. Though all of them

have almost same area under curve, the shapes of these curves were different. i.e., at a fixed sensitivity, ANN's and decision trees had higher specificity compared to LR. Figure 6.11 depicts this fact.

In summary, it is hard to draw general conclusions regarding the performance or superiority of one model over the other on the basis of findings presented in this chapter or elsewhere, since the results for each of these studies are based on the specific kind of interest. Each model has its advantages, and the selection of a model should be based on these advantages and the intended purpose of the study. In this study, we conclude that ANN model-4 and decision tree model-4 has a better predictive probability compared to logistic model and can be used as the best for the modeling and prediction of breast cancer survival.

Well-performing ANN models can be used for predictions when there is an unknown nonlinear relation between the independent variables and the dependent variables that is not well understood by other tools like logistic regression.

CHAPTER SEVEN

Conclusion and Future work

In this chapter we shall pose some possible extensions of the present research. This chapter stands on the foundations built on Chapters 2, 3, 4, 5 and 6. We make necessary connections between the methods employed in those chapters and report on ongoing work that could not be included in this thesis.

In chapter two, we have used the Inverse Gaussian (IG) distribution for statistical modeling of the breast cancer tumor sizes for the three race women. At the end of this chapter grouping ages into groups of 5, we also stratified the number of women diagnosed with breast cancer in different stages. As a future research, it would be of interest to develop a statistical/mathematical model that identifies categorized age as the independent variable and tumor size as the response variable. Having established such models which may be non-linear in nature with a high degree of accuracy, namely, high R^2 and adjusted R^2 we can further proceed to calculate the rate of change of tumor size along with age.

As a part of future research we plan to focus on the use of the kernel density estimation method. In case if we do not have enough information to fit the probability distribution of the parameters which behave as a random variable, we can proceed to investigate the applicability of the kernel density estimation method to obtain the density function of the parameters.

In chapters three and four, we have used the family of generalized extreme value distribution and log-logistic models to statistically model the survival of breast cancer women utilizing the available predictor variables for predictive purposes. As the part of future research, if we are given with more relevant information on breast cancer such as family history, age at first live birth, drinking and smoking habits, etc. we can be able to provide more compressive understanding of breast cancer. As a matter of fact, with the increased number of highly attributable variables and very handy software programs, it is of paramount importance that a survival model, incorporating such covariates be developed for more accurate and appropriate results of prediction.

For the problem of breast cancer stage classification and classification of survival or otherwise of breast cancer woman we proposed artificial neural network approach in chapters five and six. There are many areas of research that can be explored further based on the findings from these chapters. Some specific ideas for future research are listed below:

- The neural network parameters needs a random initializations of weights and biases Failing to declare proper initial values can in turn reduce the chances of proper training of network. Our proposal is to identify a relation, if any, that define neural network parameters such as weights and biases in terms of regression coefficients in statistical modeling.
- Try to identify and explore the black box nature of ANNs.
- Examine other network parameters that influence ANN performance, such as the activation function, number of hidden layers, number of epochs, learning rate, etc.

- To identify if there is any relationship between number of hidden layers, number of hidden units in each hidden layer and the output function of the network. This can mainly help us to reduce the training time.
- Evaluate the application of different activation functions mathematically and statistically to identify their ability to provide robust results.
- Linking ANNs, regression, differential equations and implementing them in applications of biological systems.

Finally, the methods used in the current study could be implemented in the study of other types of cancers in providing important information on treatment and survival of cancer patients.

REFERENCES

1. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747-752.
2. Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K., Morrison, J., Richesson, D. A., & Kropp, S. (2008). Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS genetics*, *4*(4), e1000054.
3. Colditz, G. A., Rosner, B. A., Chen, W. Y., Holmes, M. D., & Hankinson, S. E. (2004). Risk factors for breast cancer according to estrogen and progesterone receptor status. *Journal of the National Cancer Institute*, *96*(3), 218-228.
4. Berg, W. A., Zhang, Z., Lehrer, D., Jong, R. A., Pisano, E. D., Barr, R. G., ... & ACRIN 6666 Investigators. (2012). Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA*, *307*(13), 1394-1404.
5. Jemal, A., Thun, M. J., Ries, L. A., Howe, H. L., Weir, H. K., Center, M. M., ... & Edwards, B. K. (2008). Annual report to the nation on the status of cancer, 1975–2005, featuring trends in lung cancer, tobacco use, and tobacco control. *Journal of the National Cancer Institute*, *100*(23), 1672-1694.

6. Tarver, T. (2012). Cancer Facts & Figures 2012. American Cancer Society (ACS) Atlanta, GA: American Cancer Society, 2012. 66 p., pdf. Available from. *Journal of Consumer Health on the Internet*, *16*(3), 366-367.
7. Siegel, R., Naishadham, D., & Jemal, A. (2012). Cancer statistics, 2012. *CA: a cancer journal for clinicians*, *62*(1), 10-29.
8. Anderson, L. A., Pfeiffer, R., Warren, J. L., Landgren, O., Gadalla, S., Berndt, S. I., & Engels, E. A. (2008). Hematopoietic malignancies associated with viral and alcoholic hepatitis. *Cancer Epidemiology Biomarkers & Prevention*, *17*(11), 3069-3075.
9. Anderson, L. A., Landgren, O., & Engels, E. A. (2009). Common community acquired infections and subsequent risk of chronic lymphocytic leukemia. *British journal of haematology*, *147*(4), 444-449.
10. Anderson, L. A., Gadalla, S., Morton, L. M., Landgren, O., Pfeiffer, R., Warren, J. L., & Engels, E. A. (2009). Population-based study of autoimmune conditions and the risk of specific lymphoid malignancies. *International Journal of Cancer*, *125*(2), 398-405.
11. Anderson, L. A., Pfeiffer, R. M., Landgren, O., Gadalla, S., Berndt, S. I., & Engels, E. A. (2009). Risks of myeloid malignancies in patients with autoimmune conditions. *British journal of cancer*, *100*(5), 822-828.
12. Quinlan, S. C., Morton, L. M., Pfeiffer, R. M., Anderson, L. A., Landgren, O., Warren, J. L., & Engels, E. A. (2010). Increased risk for lymphoid and myeloid neoplasms in elderly solid-organ transplant recipients. *Cancer Epidemiology Biomarkers & Prevention*, *19*(5), 1229-1237.
13. Chang, C. M., Quinlan, S. C., Warren, J. L., & Engels, E. A. (2010). Blood transfusions and the subsequent risk of hematologic malignancies. *Transfusion*, *50*(10), 2249-2257.

14. Lanoy, E., & Engels, E. A. (2010). Skin cancers associated with autoimmune conditions among elderly adults. *British journal of cancer*, *103*(1), 112-114.

15. SEER-9. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence-SEER 9 Regs Research Data, Nov 2009 Sub (1973–2008) Katrinia/Rita Population Adjustment> —Linked To County Attributes—Total U.S., 1969–2007 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2011, based on November 2010 submission. 2011. http://seer.cancer.gov/.Accessed May 20, 2011.

16. SEER-13. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence-SEER 13 Regs Research Data, Nov 2010 Sub (1992–2008) —Linked To County Attributes—Total U.S., 1969–2009 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2011, based on November 2010 submission. 2011.

17. Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (McGraw-Hill, New York).

18. Portney, L. G., & Watkins, M. P. Foundations of clinical research: applications to practice, 1993. *Appleton & Lange, East Norwalk*, 148.

19. Mattson, D. E. (1981). Statistics: Difficult Concepts. *Understandable Explanations (CV Mosby, St. Louis, 1981)*.

20. Colton, T. Statistics in medicine, 1974. *Little, Brown, Boston*, *164*.

21. Gjerde, T., Eidsvik, J., Nyrnes, E., & Bruun, B. T. Normal Inverse Gaussian Error Distributions Applied for the Positioning of Petroleum Wells.

22. Chhikara, R. (1988). *The Inverse Gaussian distribution: Theory: Methodology, and Applications* (Vol. 95). CRC Press.

23. Folks, J. L., & Chhikara, R. S. (1978). The inverse Gaussian distribution and its statistical application--a review. *Journal of the Royal Statistical Society. Series B (Methodological)*, 263-289.

24. El Saghir, N. S., Seoud, M., Khalil, M. K., Charafeddine, M., Salem, Z. K., Geara, F. B., & Shamseddine, A. I. (2006). Effects of young age at presentation on survival in breast cancer. *BMC cancer*, *6*(1), 194.

25. Shannon, C., & Smith, I. E. (2003). Breast cancer in adolescents and young women. *European Journal of cancer*, *39*(18), 2632-2642.

26. Anders, C. K., Hsu, D. S., Broadwater, G., Acharya, C. R., Foekens, J. A., Zhang, Y., ... & Blackwell, K. L. (2008). Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *Journal of Clinical Oncology*, *26*(20), 3324-3330.

27. American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

28. Akritas, M. G. (2004). Nonparametric survival analysis. *Statistical Science*, 19(4), 615-623.

29. Greenwood, M. (1926). The natural duration of cancer: reports of public health and medical subjects, 33. *London: Her Majesty's Stationery Office*.

30. Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, *83*(402), 414-425.

31. Pepe, M. S., & Mori, M. (1993). Kaplan—Meier, marginal or conditional probability curves in summarizing competing risks failure time data. *Statistics in medicine*, *12*(8), 737-751.

32. Beirlant, J., & Matthys, G. (2006). Generalized extreme value distribution. *Encyclopedia of Environmetrics*.

33. Hosking, J. R. M., Wallis, J. R., & Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, *27*(3), 251-261.

34. Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal Distributions across the Sciences: Keys and Clues On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. *Bioscience*, *51*(5), 341-352.

35. Aitchison, J., & Brown, J. A. C. (1969). *The lognormal distribution, with special reference to its uses in economics* (Vol. 5). CUP Archive

36. Collett, D. (2003). *Modelling survival data in medical research*. CRC press.

37. Smith, T., Smith, B., & Ryan, M. A. (2003). Survival analysis using Cox proportional hazards modeling for single and multiple event time data. *Proceedings of the twenty-eighth annual SAS users group international conference, SAS Institute, Inc, Cary, paper* (pp. 254-28).

38. Jennison, C., & Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. CRC Press.

39. Cantor, A. (2003). *SAS survival analysis techniques for medical research*. SAS Institute.

40. Allison, P. D. Survival analysis using the SAS system: A practical guide. 1995. Cary, NC: SAS Institute.
41. Chen, Q., Oppenheim, A., & Wang, D. Survival Analysis and Data Mining.
42. Cleves, M., Gould, W., & Gutierrez, R. (2008). *An introduction to survival analysis using Stata*. Stata Press.
43. Blossfeld, H. P., & Rohwer, G. (2002). *Techniques of event history modeling: New approaches to causal analysis.* Lawrence Erlbaum Associates Publishers.
44. González-Manteiga, W., & Cadarso-Suarez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Communications in Statistics-Theory and Methods*, *4*(1), 65-78.
45. Rodríguez, G. (1994). Statistical issues in the analysis of reproductive histories using hazard models. *Annals of the New York academy of sciences*, *709*(1), 266-279.
46. Andersen, P. K., Borgan, Ø, Hjort, N. L., Arjas, E., Stene, J., & Aalen, O. (1985). Counting process models for life history data: A review [with discussion and reply]. *Scandinavian Journal of Statistics*, 97-158.
47. Buis, M. L. (2006). An introduction to survival analysis. *Department of Social Research Methodology Vrije Universiteit Amsterdam [Online]*.
48. Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data: Using SAS*. SAS Institute.
49. Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.
50. Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (Vol. 360). John Wiley & Sons.

51. Gardiner, J. C., Luo, Z., Bradley, C. J., Sirbu, C. M., & Given, C. W. (2006). A dynamic model for estimating changes in health status and costs. *Statistics in medicine, 25*(21), 3648-3667.

52. Gray, R. J., & Pierce, D. A. (1985). Goodness-of-fit tests for censored survival data. *The Annals of Statistics*, 552-563.

53. Aaserud, S., Kvaløy, J. T., & Lindqvist, B. H. (2013). Residuals and functional form in accelerated life regression models. In *Risk Assessment and Evaluation of Predictions* (pp. 61-65). Springer New York.

54. Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal Distributions across the Sciences: Keys and Clues On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. *Bioscience, 51*(5), 341-352.

55. Bennett, S. (1983). Log-logistic regression models for survival data. *Applied Statistics*, 165-171.

56. Cox, D. R. (1972). Regression models and life tables. *JR stat soc B, 34*(2), 187-220.

57. Cox, D. R. (1975). Partial likelihood. *Biometrika, 62*(2), 269-276.

58. Klein, J. P., & Zhang, M. J. (2005). *Survival analysis, software*. John Wiley & Sons, Ltd.

59. Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *IEEE computer, 29*(3), 31-44.

60. Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural networks. *Cancer, 91*(S8), 1615-1635.

61. Looney, Carl Grant. *Pattern recognition using neural networks: theory and algorithms for engineers and scientists*. Oxford University Press, Inc., 1997.

62. Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(1), 30-42.

63. Al-Shayea, Q. K. (2011). Artificial Neural Networks in Medical Diagnosis. *International Journal of Computer Science Issues (IJCSI)*, *8*(2).

64. Mano, M., Capi, G., Tanaka, N., & Kawahara, S. (2013). An artificial neural network based robot controller that uses rat's brain signals. *Robotics*, *2*(2), 54-65.

65. Cochocki, A., & Unbehauen, R. (1993). *Neural networks for optimization and signal processing*. John Wiley & Sons, Inc.

66. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

67. Minsky, M., & Seymour, P. (1969). Perceptrons.

68. Indiveri, G., Linares-Barranco, B., Hamilton, T. J., Van Schaik, A., Etienne-Cummings, R., Delbruck, T., & Boahen, K. (2011). Neuromorphic silicon neuron circuits. *Frontiers in neuroscience*, *5*.

69. Lippmann, R. P. (1987). An introduction to computing with neural nets. *ASSP Magazine, IEEE*, *4*(2), 4-22.

70. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), 113-127.

71. Park, J. H., Kim, Y. S., Eom, I. K., & Lee, K. Y. (1993). Economic load dispatch for piecewise quadratic cost function using Hopfield neural network. *Power Systems, IEEE Transactions on*, *8*(3), 1030-1038.

72. Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. *Cambridge, MA: MIT Press*, *1*, 282-317.

73. Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural networks*, *10*(9), 1659-1671.

74. Mitra, S., & Hayashi, Y. (2000). Neuro-fuzzy rule generation: survey in soft computing framework. *Neural Networks, IEEE Transactions on*, *11*(3), 748-768.

75. Bishop, C. M., Svensén, M., & Williams, C. K. (1998). GTM: The generative topographic mapping. *Neural computation*, *10*(1), 215-234.

76. Börgers, T., & Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, *77*(1), 1-14.

77. Fritzke, B. (1994). Growing cell structures—a self-organizing network for unsupervised and supervised learning. *Neural networks*, *7*(9), 1441-1460.

78. Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feed forward neural network. *Neural networks*, *2*(6), 459-473.

79. Ruck, D. W., Rogers, S. K., Kabrisky, M., Oxley, M. E., & Suter, B. W. (1990). The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *Neural Networks, IEEE Transactions on*, *1*(4), 296-298.

80. Nigrin, A. (1993). *Neural networks for pattern recognition*. MIT Press.

81. Ripley, B. D. (1996). Pattern recognition via neural networks. *A volume of Oxford Graduate Lectures on Neural Networks, title to be decided. Oxford University Press. [See http: //www. stats. ox. ac. uk/ripley/papers. html.]*.

82. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303-314.

83. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feed forward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

84. Hartman, E. J., Keeler, J. D., & Kowalski, J. M. (1990). Layered neural networks with Gaussian hidden units as universal approximations. *Neural computation*, *2*(2), 210-215.

85. Pao, Y. H., & Sobajic, D. J. (1992). Combined use of unsupervised and supervised learning for dynamic security assessment. *Power Systems, IEEE Transactions on*, *7*(2), 878-884.

86. Leung, H., & Haykin, S. (1993). Rational function neural network. *Neural Computation*, *5*(6), 928-938.

87. Dorffner, G. (1994). Unified framework for MLPs and RBFNs: Introducing conic section function networks. *Cybernetics and Systems: An International Journal*, 25(4), 511-554.

88. Giraud, B. G., Lapedes, A., Liu, L. C., & Lemm, J. C. (1995). Lorentzian neural nets. *Neural Networks*, *8*(5), 757-767.

89. Duch, W., & Jankowski, N. (2001, April). Transfer functions: hidden possibilities for better neural networks. In *ESANN* (pp. 81-94).

90. Duch, W., & Jankowski, N. (1999). Survey of neural transfer functions. *Neural Computing Surveys*, *2*(1), 163-212.

91. Karlik, B., & Olgac, A. V. (2010). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, *1*(4), 111-122.

92. Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*(5), 654-657.

93. van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530-536.

94. Poornashankar. "Performance Analysis of Different Feed Forward Networks in Non-Linear Classification." International Journal of Soft Computing & Engineering 3.2:332.

95. Hartman, E., & Keeler, J. D. (1991). Predicting the future: Advantages of semi local units. *Neural Computation*, *3*(4), 566-578.

96. *Performance Comparison of Different Multilayer Perceptron Network Activation Functions in Automated Weather Classification.* I. S. Isa, S.Omar, Z. Saad, M. K. Osman**.** Kota Kinabalu, Malaysia: Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation (AMS), 2010, pp. 71-75. 978-1-4244-7196-6.

97. Agresti, A. (2010). *Analysis of ordinal categorical data* (Vol. 656). John Wiley & Sons.

98. Tu, J. V., & Guerriere, M. R. (1992). Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (p. 666). American Medical Informatics Association.

99. Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

100. McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

101. Verhulst, P. F. (1977). A note on the law of population growth. In *Mathematical Demography* (pp. 333-339). Springer Berlin Heidelberg.

102. Cramer, J. S. (2003). The origins and development of the logit model. *Logit models from economics and other fields*, 149-158.

103. Reed, L. J., & Berkson, J. (1929). The application of the logistic function to experimental data. *The Journal of Physical Chemistry*, *33*(5), 760-779.

104. Yule, G. U. (1919). *An introduction to the theory of statistics*. C. Griffin, limited.

105. Berkson, Joseph. "Why I prefer logits to probits." *Biometrics* 7.4 (1951): 327-339.

106. McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology*, *4*(1), 103-120.

107. Trexler, J. C., & Travis, J. (1993). Nontraditional regression analyses. *Ecology*, 1629-1637.

108. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

109. Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965-980.

110. Friedman, J., Hastie, T., & Tibshirani, R. (2000). Special invited paper. Additive logistic regression: A statistical view of boosting. *Annals of statistics*, 337-374.

111. Hartman, R. S. (1988). Self-selection bias in the evolution of voluntary energy conservation programs. *The Review of Economics and Statistics,* 448-458.

112. Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

113. Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)--a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.

114. Widrow, B., & Lehr, M. A. (1990). 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, *78*(9), 1415-1442.

115. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). *Learning representations by back-propagating errors* (pp. 696-699). MIT Press, Cambridge, MA, USA.

116. Jabri, M., & Flower, B. (1992). Weight perturbation: An optimal architecture and learning technique for analog VLSI feedforward and recurrent multilayer .networks. *Neural Networks, IEEE Transactions on*, *3*(1), 154-157.

117. Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, *6*(4), 525-533.

118. Mohr, A. E., Riskin, E. A., & Ladner, R. E. (2000). Unequal loss protection: Graceful degradation of image quality over packet erasure channels through forward error correction. *Selected Areas in Communications, IEEE Journal on*, *18*(6), 819-828.

119. Deeks, J. J., & Altman, D. G. (2004). Statistics Notes: Diagnostic tests 4: likelihood ratios. *BMJ: British Medical Journal*, *329*(7458), 168.

120. Banks, E. (2001). Hormone replacement therapy and the sensitivity and specificity of breast cancer screening: a review. *Journal of medical screening*, *8*(1), 29-35.

121. Barlow, W. E., White, E., Ballard-Barbash, R., Vacek, P. M., Titus-Ernstoff, L., Carney, P. A., & Kerlikowske, K. (2006). Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, *98*(17), 1204-1214.

122. Terris, M. K., Wallen, E. M., & Stamey, T. A. (1997). Comparison of mid-lobe versus lateral systematic sextant biopsies in the detection of prostate cancer.*Urologia internationalis*, *59*(4), 239-242.

123. Bruix, J., & Llovet, J. M. (2002). Prognostic prediction and treatment strategy in hepatocellular carcinoma. *Hepatology*, *35*(3), 519-524.

124. Lisboa, P. J., & Taktak, A. F. (2006). The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, *19*(4), 408-415.

125. Baxt, W. G. (1995). Application of artificial neural networks to clinical medicine. *The lancet*, *346*(8983), 1135-1138.

126. International Bladder Cancer Nomogram Consortium. (2006). Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer. *Journal of Clinical Oncology*, *24*(24), 3967-3972.

127. Mermelstein, R. J., & Riesenberg, L. A. (1992). Changing knowledge and attitudes about skin cancer risk factors in adolescents. *Health Psychology*, *11*(6), 371.

128. Rodríguez, G. (2007). Lecture Notes on Generalized Linear Models. Available at http://data.princeton.edu/wws509/notes/

129. Karabatak. M. and M. Cevdet. An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications 36: 3465–3469, 2009.

130. Kovalerchuk, B., Triantaphyllou, E., Ruiz, J. F., & Clayton, J. (1997). Fuzzy logic in computer-aided breast cancer diagnosis: analysis of lobulation. Artificial Intelligence in Medicine, 11(1), 75-85.
131. Sujatha, G., & Rani, K. U. An Experimental Study on Ensemble of Decision Tree Classifiers.
132. Quinlan, J. R. (1987). Simplifying decision trees. International journal of man-machine studies, 27(3), 221-234.
133. Zantema, H., & Bodlaender, H. L. (2000). Finding small equivalent decision trees is hard. International Journal of Foundations of Computer Science, 11(02), 343-354.
134. Friedman, J. H., Kohavi, R., & Yun, Y. (1996, August). Lazy decision trees. InAAAI/IAAI, Vol. 1 (pp. 717-724).
135. Rokach, L., & Maimon, O. (2005). Decision trees. In Data Mining and Knowledge Discovery Handbook (pp. 165-192). Springer US.
136. Olshen, L. B. J. F. R., & Stone, C. J. (1984). Classification and regression trees. Wadsworth International Group.
137. Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons.
138. Friedman, J. H. (1977). A recursive partitioning decision rule for nonparametric classification. IEEE Trans. Computers, 26(4), 404-408.
139. Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. Machine learning, 4(2), 227-243.

140. Last, M., Maimon, O., & Minkov, E. (2002). Improving stability of decision trees.International Journal of Pattern Recognition and Artificial Intelligence, 16(02), 145-159.

141. Janikow, C. Z. (1998). Fuzzy decision trees: issues and methods. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 28(1), 1-14.

142. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.