5-27-2014

Statistical Analysis, Modeling, and Algorithms for Pharmaceutical and Cancer Systems

Bong-Jin Choi

University of South Florida, bchoi@mail.usf.edu

Follow this and additional works at: https://scholarcommons.usf.edu/etd

Part of the Medicinal Chemistry and Pharmaceutics Commons, Medicine and Health Sciences Commons, and the Statistics and Probability Commons

Scholar Commons Citation

https://scholarcommons.usf.edu/etd/5200

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Statistical Analysis, Modeling, and Algorithms for Pharmaceutical and Cancer Systems

by

Bong-Jin Choi

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

Major Professor: Chris P. Tsokos, Ph.D
Marcus McWaters, Ph.D.
Gangaram Ladde, Ph.D.
Kandethody M. Ramachandran, Ph.D.
Jongphil Kim, Ph.D.

Date of Approval:
May 27, 2014

Keywords: Survival Analysis, Decision Tree, Random Forest, Variable Rank, Drug Efficiency

Copyright © 2014, Bong-Jin Choi
Dedication

To My Parents, Hayeon, and Junghyun.
Acknowledgements

I would like to thank everyone helped and inspired me during my graduate studies at USF. I would firstly like to express my sincerest and deepest gratitude to my advisor, Distinguished University Professor Chris P. Tsokos. His diligence, critical thinking, and genuine care and concern for students have greatly inspired me. His constant aid and advice throughout the entire process perpetuated my interest in research. I am especially grateful for his patience, encouragement, and faith in me.

I am also truly grateful for Professors Marcus Mcwaters, Gangarm Ladde, and Kandethody Ramachandran for their guidance and kind assistance during my dissertation research. I am also indebted to Dr. Ji-Hyun Lee and Dr. Chanyoung Lee for their support and assistance.

I would also like to thank H. Lee Moffitt Cancer Center and Research Institute, Center for Urban Transportation Research, and especially the Department of Mathematics and Statistics for providing financial assistance to pursue my doctoral degree.

Thanks to all my good friends, Daniel Siu, Dooyoung Kim, Carlos, Chunlin Cong, Young Xu, and my other fellow graduate students for the beautiful moments we shared together.
# Table of Contents

List of Tables .................................................................................................................. iii
List of Figures .................................................................................................................. v
Abstract .............................................................................................................................. vii

Chapter 1 Introduction ....................................................................................................... 1
  1.1 A Statistical Algorithm For Determining The Optimal Doses Of Drugs: 
    A Five Parameters Logistic Model ........................................................................... 1
  1.2 Survival Analysis of Breast Cancer Data Using the Random Forests: an 
    Ensemble of Trees ................................................................................................. 2
  1.3 Estimating the Four Parameters Johnson-SB Probability Distribution: 
    New R Package ...................................................................................................... 3
  1.4 Estimating Survival Time of Baseline Lung Cancer Using Statistical 
    Modeling .................................................................................................................. 4

Chapter 2 A Statistical Algorithm For Determining The Optimal Doses Of Drugs . 5
  2.1 Introduction ............................................................................................................ 5
  2.2 The Four Parameter Logistic Model ..................................................................... 6
  2.3 The Five Parameter Logistic Model ..................................................................... 9
  2.4 The Effects of The Five Parameters In The Logistic Model ............................... 10
  2.5 Development Of A Statistical Algorithms To Estimate The Five Parameters 
    Logistic Model ...................................................................................................... 13
  2.6 Drug Efficiency Estimator Program .................................................................... 17
  2.7 Compare with Commercial Program .................................................................... 18
  2.8 Summary and Contribution .................................................................................. 19

Chapter 3 Survival Analysis of Breast Cancer Data Using the Random Forests: an 
  Ensemble of Trees ...................................................................................................... 21
  3.1 Introduction ............................................................................................................ 21
  3.1.1 Introduction to Decision Tree ......................................................................... 22
  3.1.2 Theory Behind Decision Tree Analysis .......................................................... 23
  3.1.3 Random Forest ............................................................................................... 25
  3.2 About Data ............................................................................................................ 27
  3.3 Data pre-processing Methods .............................................................................. 30
  3.4 Pre-Processing ...................................................................................................... 31
  3.5 Validation of Procedure ....................................................................................... 36
  3.6 Summary and Contribution .................................................................................. 37
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Start Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>39</td>
</tr>
<tr>
<td>4.2</td>
<td>Numerical Methods</td>
<td>40</td>
</tr>
<tr>
<td>4.3</td>
<td>How to Use the JohnsonUSF Package</td>
<td>44</td>
</tr>
<tr>
<td>4.4</td>
<td>Validation of Estimated Parameters</td>
<td>45</td>
</tr>
<tr>
<td>4.5</td>
<td>Apply to Real Lung Cancer Data</td>
<td>47</td>
</tr>
<tr>
<td>4.6</td>
<td>Conclusion</td>
<td>47</td>
</tr>
<tr>
<td>4.7</td>
<td>Contribution</td>
<td>49</td>
</tr>
<tr>
<td>5</td>
<td>Estimating Survival Time of Baseline Lung Cancer Using Statistical Modeling</td>
<td>50</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>50</td>
</tr>
<tr>
<td>5.2</td>
<td>Lung Cancer Data with Smoking Information</td>
<td>51</td>
</tr>
<tr>
<td>5.3</td>
<td>Classified Groups</td>
<td>52</td>
</tr>
<tr>
<td>5.4</td>
<td>Correlation Matrix</td>
<td>52</td>
</tr>
<tr>
<td>5.5</td>
<td>Survival Analysis and Statistical Modeling</td>
<td>55</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Multiple Regression Model for Group I</td>
<td>56</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Multiple Regression Model for Group II</td>
<td>58</td>
</tr>
<tr>
<td>5.5.3</td>
<td>Multiple Regression Model for Group III</td>
<td>60</td>
</tr>
<tr>
<td>5.5.4</td>
<td>Multiple Regression Model for Group IV</td>
<td>62</td>
</tr>
<tr>
<td>5.6</td>
<td>Fitting the Best Possible Probability Distribution</td>
<td>64</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Goodness of Fit Test for Group All</td>
<td>64</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Goodness of Fit Tests for Group I</td>
<td>66</td>
</tr>
<tr>
<td>5.6.3</td>
<td>Goodness of Fit Test for Group II</td>
<td>67</td>
</tr>
<tr>
<td>5.6.4</td>
<td>Goodness of Fit Test for Group III</td>
<td>69</td>
</tr>
<tr>
<td>5.6.5</td>
<td>Goodness of Fit Test for Group IV</td>
<td>71</td>
</tr>
<tr>
<td>5.7</td>
<td>Conclusion and Contributions</td>
<td>74</td>
</tr>
<tr>
<td>6</td>
<td>Future Research Projects</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>About the Author</td>
<td>End Page</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Description of Discrete Variables.</td>
<td>32</td>
</tr>
<tr>
<td>Table 2</td>
<td>Basic Summary of Discrete Variables.</td>
<td>32</td>
</tr>
<tr>
<td>Table 3</td>
<td>Rank for Risk Factors.</td>
<td>34</td>
</tr>
<tr>
<td>Table 4</td>
<td>Summary of Rank for Risk Factors.</td>
<td>34</td>
</tr>
<tr>
<td>Table 5</td>
<td>Log Rank Test.</td>
<td>37</td>
</tr>
<tr>
<td>Table 6</td>
<td>Functions in JohnsonUSF Package.</td>
<td>45</td>
</tr>
<tr>
<td>Table 7</td>
<td>Basic Statistics for the Parameters $\delta$ and $\gamma$ for Ten Thousand Johnson SB Probability Distributed Random Samples.</td>
<td>45</td>
</tr>
<tr>
<td>Table 8</td>
<td>Kolmogorov-Smirnov, Anderson-Darling, and Chi-Square goodness of fit test results.</td>
<td>46</td>
</tr>
<tr>
<td>Table 9</td>
<td>Estimate Comparison Using JohnsonUSF Package and Other Software Program.</td>
<td>47</td>
</tr>
<tr>
<td>Table 10</td>
<td>Correlation Matrix on Caucasian, Current Smoker, and Male (Group I).</td>
<td>52</td>
</tr>
<tr>
<td>Table 11</td>
<td>Correlation Matrix on Caucasian, Current Smoker, and Female (Group II).</td>
<td>53</td>
</tr>
<tr>
<td>Table 12</td>
<td>Correlation Matrix White, Former Smoker, and Male (III).</td>
<td>53</td>
</tr>
<tr>
<td>Table 13</td>
<td>Correlation Matrix on Caucasian, Former Smoker, and Female (IV).</td>
<td>54</td>
</tr>
<tr>
<td>Table 14</td>
<td>Variables Description.</td>
<td>55</td>
</tr>
<tr>
<td>Table 15</td>
<td>Residual Validation : Group I (Mean residual : 0.075 and standard error : 0.95).</td>
<td>57</td>
</tr>
<tr>
<td>Table 16</td>
<td>Residual Validation : Group II (Mean residual : -0.023 and standard error : 0.94).</td>
<td>59</td>
</tr>
<tr>
<td>Table 17</td>
<td>Residual Validation : Group III (Mean residual : -0.158 and standard error : 1.06).</td>
<td>61</td>
</tr>
<tr>
<td>Table 18</td>
<td>Residual Validation : Group IV (Mean residual : -0.449 and standard error : 0.993).</td>
<td>63</td>
</tr>
<tr>
<td>Table 19</td>
<td>Summary of Goodness-of-fit Test for Group All.</td>
<td>64</td>
</tr>
</tbody>
</table>
Table 20  Summary Of Goodness-of-fit Test For Group I. .......................... 66
Table 21  Summary of Goodness-of-fit test for Group II. .......................... 68
Table 22  Summary of Goodness-of-fit Test For Group III. ........................ 70
Table 23  Summary of Goodness-of-fit Test For Group IV. ........................ 72
Table 24  Summary of Mathematical Models, Probability Density and Survival Function. ......................................................... 75
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Four Parameter Logistic Model with A=5, B=5.5, C=5, D=5</td>
<td>7</td>
</tr>
<tr>
<td>Figure 2</td>
<td>The Effect of Parameters A in five parameters logistic model based on ( y = D + \frac{A-D}{(1+\frac{x}{c})^B} )</td>
<td>11</td>
</tr>
<tr>
<td>Figure 3</td>
<td>The Effect of Parameters B in five parameters logistic model based on ( y = D + \frac{A-D}{(1+\frac{x}{c})^B} )</td>
<td>11</td>
</tr>
<tr>
<td>Figure 4</td>
<td>The Effect of Parameters C in five parameters logistic model based on ( y = D + \frac{A-D}{(1+\frac{x}{c})^B} )</td>
<td>12</td>
</tr>
<tr>
<td>Figure 5</td>
<td>The Effect of Parameters D in five parameters logistic model based on ( y = D + \frac{A-D}{(1+\frac{x}{c})^B} )</td>
<td>12</td>
</tr>
<tr>
<td>Figure 6</td>
<td>The Effect of Parameters g in five parameters logistic model based on ( y = D + \frac{A-D}{(1+\frac{x}{c})^B} )</td>
<td>13</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Main Modules for Drug Efficiency Program.</td>
<td>14</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Dynamic Initial Value Detection.</td>
<td>15</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Algorithm of Drug Efficiency Estimate Program.</td>
<td>16</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Main Figure of Drug Efficiency Estimate Program.</td>
<td>17</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Result of Drug Efficiency Estimate Program.</td>
<td>18</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Basic Decision Tree.</td>
<td>23</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Systematic Diagram To Build Random Forest.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 14</td>
<td>SEER Breast Cancer Diagram.</td>
<td>28</td>
</tr>
<tr>
<td>Figure 15</td>
<td>SEER Breast Cancer Histological Diagram.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Normal Breast with Invasive Ductal Cancer : from <a href="http://www.breastcancer.org">www.breastcancer.org</a>.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 17</td>
<td>SEER Breast Cancer Diagram : Caucasians.</td>
<td>30</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Step-by-Step Procedure to Remove Non Significant Variables.</td>
<td>31</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Program Algorithm Flow Chart.</td>
<td>33</td>
</tr>
</tbody>
</table>
Figure 20 Final Tree. ................................................................. 35
Figure 21 Survival Curves using Final Classified Groups. .................. 35
Figure 22 Johnson’s SB distribution with (γ, δ) combinations first row-left(0.0.5)
  , first row-middle(0.0.75), first row-right (0.25,0.5), second row-
  left(0.5,0.5),second row-middle (1,2),second row-right (1,1). ........ 41
Figure 23 Algorithm to estimate the four parameters of the Johnson SB Prob-
  ability Distribution. ............................................................. 43
Figure 24 Using local JohnsonUSF.zip file to install on your computer. ...... 44
Figure 25 Random Sample from JohnsonSB Probability Distribution
  δ = 3, γ = 4, ξ = 100, λ = 0. ...................................................... 46
Figure 26 Data Fitting Using Real Lung Cancer Data. .......................... 48
Figure 27 Data Diagram of Lung Cancer’s Patients Information. .............. 51
Figure 28 Matrix of Scatter Plots for All Groups. .............................. 54
Figure 29 Matrix of Scatter Plots for Groups1. .................................. 56
Figure 30 Matrix of Scatter Plots for Groups2. .................................. 58
Figure 31 Matrix of Scatter Plots for Groups3. .................................. 60
Figure 32 Matrix of Scatter Plots for Groups4. .................................. 62
Figure 33 Compare Survival Function with Parametric Method and K-M in
  Group All. ................................................................. 65
Figure 34 Compare Survival Function with Parametric Method and K-M in
  Group I. ................................................................. 67
Figure 35 Compare Survival Function with Parametric Method and K-M in
  Group II. ................................................................. 69
Figure 36 Compare Survival Function with Parametric Method and K-M in
  Group III. ................................................................. 71
Figure 37 Compare Survival Function with Parametric Method and K-M in
  Group IV. ................................................................. 73
Figure 38 Compare Survival Curve Between Groups And Overall Survival
  Time. ................................................................. 73
Figure 39 Estimated Survival Time and Probability with Three factors. ....... 76
Abstract

The goal of this PhD research project is to devise a robust interatomic potential for large-scale molecular dynamics simulations of carbon materials under extreme conditions. This screened-environment dependent reactive empirical bond order potential (SED-REBO) is specifically designed to describe carbon materials under extreme compressive or tensile stresses. Based on the original REBO potential by Brenner and co-workers, SED-REBO includes reparametrized pairwise interaction terms and a new screening term, which serves the role of a variable cutoff. The SED-REBO potential overcomes the deficiencies found with the most commonly used interatomic potentials for carbon: the appearance of artificial forces due to short cutoff that are known to create erroneous phenomena including ductile fracture of graphene and carbon nanotubes, which contradicts the experimentally observed brittle character of these materials. SED-REBO was applied in large scale molecular dynamics simulations of nanoindentation of graphene membranes and shock-induced compression of diamond. It was shown in the first computational experiment that graphene membranes exhibit a non-linear response to large magnitude of indentation, followed by a brittle fracture in agreement with experiments. The strength of graphene was determined using the kinetic theory of fracture, and the crack propagation mechanisms in the material were identified. It was found in large-scale shock simulations that SED-REBO improves the predictive power of MD simulations of carbon materials at extreme conditions.
Chapter 1
Introduction

The subject study consists of four chapters and given below are brief introductions.

1.1 A Statistical Algorithm For Determining The Optimal Doses Of Drugs: A Five Parameters Logistic Model

Drug effects are commonly screened with cell lines before animal studies and clinical trials are conducted. Using the point estimate of the half maximal inhibitory concentration (IC50) as surrogates, drug potency and cell line sensitivity are determined. In the present study, we demonstrate how the five parameters logistic(5PL) function performs better over the four parameters logistic(4PL) and its variants. Specifically, we study the behaviors in each functions and the improvement in accuracy of concentration estimates(IC50) using the 5PL over the 4PL as a function of the asymmetry present in the data.

The goal of this chapter is to obtain concentration estimates(IC50) that are as accurate as possible. We can improve this by the quality of the fit of the data. Two of the reasons that this can be an issue lies on either pure error or lack-of-fit error. Pure error is identified as the presence of random variation in the data, and can be reduced by increasing the number of standard replicates. Lack-of-fit represents an error in the curve modeling of the data. Unlike the first error, this cannot be reduced by increasing the number of standard replicates. For example, much immunoassay data have sigmoidal, or "s", shapes, if the data are taken over a wide enough concentration ranges. If a straight line is used as the curve model to fit such data, fitting data would be the main issue simply because a straight line cannot fit the curved shape.
Bias is introduced into the dose estimates when the quality of the fit of the data is low. Accuracy is also reduced because over-parameterized curve models have a tendency to misrepresent the true curve between data points. Over-parameterized models have the extended problem that they are more susceptible to noise in the responses, which reduces the precision of dose estimates.

1.2 Survival Analysis of Breast Cancer Data Using the Random Forests: an Ensemble of Trees

Tree-based methods by Bell(1999),[2] have become popular in performing survival analysis with complex data structures. The data chosen for this study is the SEER data obtained from their website,[8]. Within the Random Forest created by Breiman (2001),[4], the decision tree analysis by Yuan(1995),[32] was applied to identify the most important attributable variables which are the major influence in the survival time estimate of a given cancer patient. Statistically, this information is then extended to develop a statistical model estimating survival time. This approach proposes to reduce subject estimate prediction errors using R and My-SQL.

Data mining techniques by Han(2000),[14], have been extensively applied to breast cancer diagnosis, predicting the presence of cancer and differentiating between malignant and benign cases.

WHO statistical reports by D.M. Parkin (2005),[6] report that the incidence of breast cancer is the number one form of cancer among women. Breast cancer occurs due to an uncontrolled growth of cells in the breast tissues by BCF,[3] and damage to the cell’s DNA, National Breast Cancer Center,[22]. A tumor is an abnormal cell growth that can either be benign or malignant. Benign tumors are non invasive while malignant tumors are cancerous and spread to other parts of the body. In the United States, approximately one in eight women have a risk of developing breast cancer over their lifetime. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and
80% after 10 years of diagnosis, AMSWebite.[30]. Early diagnosis and treatment help to prevent the spreading of cancer.

In this chapter, we used the SEER breast cancer data and introduced the classification approach according to our research interest. Data mining was also explained as the process of analyzing large quantities of data and summarizing them into useful information. We then extended this process to create classified groups and rank attributable variables using a weighted counting method.

1.3 Estimating the Four Parameters Johnson-SB Probability Distribution: New R Package

Scientists often experience difficulties in obtaining estimates for the four parameters of the Johnson SB probability distribution. The Four parameter Johnson SB probability distribution is very important in parametrically studying a variety of health, environmental, and engineering problems.

Many methods have been proposed to estimate the parameters of the Johnson SB probability distribution, one of them being the four percentile method,Wheeler, Slifker and Shapiro, 1980. Another method to obtain approximate estimates is the Knoebel-Burkhart method (Knoebel and Burkhart,[27] at 1991)

The aim of the present study is to review such existing methods and their ability to estimate the four parameters of the subject probability density function. We are proposing a fast algorithm that produces better approximations for these parameters using numerical simulations, real cancer data, and normal transformations to generate random numbers and the Newton-Raphson method. Graphs using R are also drawn up by incorporating general integration and a prediction procedure. These methods and procedures all come together to supply the backbone in the effectiveness of our method.
1.4 Estimating Survival Time of Baseline Lung Cancer Using Statistical Modeling

Lung Cancer is the second leading cause of death and was among one of the first diseases causally linked to smoking. The relationship between smoking and lung cancer is well documented (Doll and Peto, 1978; Doll et al., 2004; IARC, 2004), along with the benefits of smoking cessation (US DHHS, 1990; Peto et al., 2000) among other. In the United States, cigarette smoking causes approximately 438,000 deaths each year, including deaths related to cancer, cardiovascular disease, respiratory disease and infant mortality [1]. Compared to nonsmokers, men who smoke are about 23 times more likely to develop lung cancer and women who smoke are about 13 times more likely. Smoking accounts for 90% of all cases, making it a major determinant of lung cancer. Many studies have focused on two hypotheses: whether the combined effect of asbestos and smoking is additive (each factor acts independently) or multiplicative (the effect of asbestos exposure on lung cancer risk is proportional to the effect of smoking) (Doll and Peto, 1985; Hammond et al., 1979; Lee, 2001). In our research, the purpose is to identify the significant attributable variables as well as all possible interactions between those variables. We also aim to construct a statistical model to predict the survival time as a function of those attributable variables and interactions. In this study, we used the number of cigarettes per day, duration of smoking, and the age at diagnosis of lung cancer to identify the significant attributable variables, including their interactions, the survival time, and probability density functions for the classified groups.
Chapter 2
A Statistical Algorithm For Determining The Optimal Doses Of Drugs

2.1 Introduction

Drug effect using cell lines is commonly screened before animal studies and clinical trials are conducted. The point estimate of the half maximal inhibitory concentration (IC50) is commonly used as the surrogate for drug potency and cell line sensitivity toward a given drug. A good method which could provide reliable parameter estimation and biological interpretability is critical for drug development. Several typical issues associated with estimating IC50 using nonlinear statistical models, such as logistic regression, include the following: control level of cell survival is not always 100% because cells proliferate and not all the cells are necessarily killed off at the highest given dose (i.e., background-level cell survival does not always reach 0%). In the present study, we demonstrate how the use of the five parameters logistic(5PL) function can improve assay performance over the four parameters logistic(4PL) and its variants. Specifically, the improvement in the accuracy of concentration estimates(IC50) that can be obtained using the 5PL over the 4PL as a function of the asymmetry present in the data is studied. The behavior of the 5PL curve and the 4PL curve are discussed.

The goal of this chapter is to obtain concentration estimates(IC50) that are as accurate as possible by improving the quality of the fit of the data. However, there are two reasons that the curve will not fit the data perfectly. The first reason is the presence of random variation in the data. This kind of error is called pure error and can be reduced by increasing the number of replicates of each standard. The second reason is that the curve model may
not approximate the true curve very well. This kind of error is called lack-of-fit error and cannot be reduce by increasing the number of standard replicates. For example, much immunoassay data have a sigmoidal, or "s", shape, if data are taken over a wide enough range of concentrations. If a straight line is used as the curve model to fit such data, much of the error in fitting the data would be due to lack-of-fit error. This is because a straight line cannot fit the curving sigmoidal shape of the data.

When the curve model is unable to follow the true curve, bias is introduced into the dose estimates and their accuracy is reduced. Furthermore, because over-parameterized curve models tend not to represent the true curve very well between data points, accuracy is reduced in these models, even though such curve models may "fit" the data quite well. Over-parameterized models have the extended problem that they are more susceptible to noise in the responses, which reduces the precision of dose estimates.

2.2 The Four Parameter Logistic Model

Logistic functions are useful for drug efficiency modeling. Among all the possible forms of logistic functions, the four-parameter logistic(4PL) function is the most commonly used in drug efficiency modeling. The parametric form of this four-parameter function is shown in equation (2.1) and graphically by Figure 1.

\[
y = D + \frac{A - D}{1 + \left(\frac{x}{C}\right)^B},
\]  

(2.1)

\[
x = f^{-1}(y; A, B, C, D)
\]

\[
= c \left( \frac{y - A}{D - y - 1} \right)^{\frac{1}{B}},
\]  

(2.2)

- A = Minimum asymptote. In a bioassay where you have a standard curve, this can be
thought of as the response value at a 0 standard concentration.

• B = Hill’s slope. The Hill’s slope refers to the steepness of the curve. This parameter controls the speed to the asymptotes. It can either be positive or negative.

• C = Inflection point. The inflection point is defined as the point on the curve where the curvature changes direction or signs.

• D = Maximum asymptote. In an bioassay where you have a standard curve, this can be thought of as the response value for infinite standard concentration.

Figure 1.: Four Parameter Logistic Model with A=5, B=5.5, C=5, D=5

To quantify the concentration of the analysis, the response must be compared to a calibration curve, commonly called the ideal curve. The unknown concentration of an analysis may then be determined by finding the concentration on the standard curve that produces the same response as that obtained from the unknown sample.

We define a true curve if and only if we have an infinite number of concentrations, each with an infinite number of replicates. Ideally, the standard curve would be identical to the
true data curve: the curve that expresses the concentration versus response relationship without any degradation by errors.

The four-parameter logistic (4PL) function is widely used in practice and is closely related to the linear logit-log model. However, like the logit-log model, the 4PL model cannot effectively model asymmetric data. The mass action model is the only physically based model, has many parameters, which makes it less able to average out noise than models with fewer parameters. It has also been shown that certain approximations to the mass action model are approximately equivalent to the logit-log model, which, as noted above, implies approximate equivalence to the 4PL curve. Like the 4PL, this approximation to the full mass action model is unable to fit asymmetric data.

A method that is used to handle asymmetry is the log 4PL method. This method takes advantage of the fact that taking the logarithm of the response of some asymmetric assay data can make the data more symmetrical and therefore better suited to a 4PL fit. This approach can improve the fit when the low response end of a sigmoidal curve has a shorter tail than the upper end of the curve.

However, taking the log of the responses of sigmoidal data where the responses on the low response end of the curve approach the lower asymptote more slowly than the responses on the high response end makes the data more asymmetric. Since this type of behavior is encountered in immuno assay and bio assay data more often than the former, this approach is not fit for routine data reduction. Many other curve models have been tried with varying degrees of success. Space does not permit further discussion here, but Rodbard provides a useful overview of some of the dose response models that have been discussed in the literature.

The 4PL function has an assumption of symmetry about the IC50 value which is rarely tested in many applications. When a curve is unable to accommodate the asymmetry, bias is introduced into the potency estimate due to lack of fit.
Recently, the five-parameter logistic (5PL) function has seen increased use as a model for bioassay dose-response curves with an extra parameter to characterize the curve asymmetry. This 5PL, as shown by equation 2.2, below, includes the 4PL function as a special case when \( G = 1 \). The 5PL function takes curve asymmetry into consideration to overcome some drawbacks of the four parameter logistic (4PL) function.

The 4PL model can be extended by adding a fifth parameter that controls the degree of asymmetry of the curve [25] [20]. This is available commercially [5]. With the extra flexibility afforded by its asymmetrical parameter, the 5PL model is able to virtually eliminate the lack-of-fit error that occurs when the 4PL is fitted to asymmetric dose response data. This chapter focuses on the use of the 5PL model for fitting dose response data. The 5PL model provides an excellent compromise between over-parameterized models that can fit data closely at the cost of a large variance in the predictions and under-parameterized models that suffer from large lack-of-fit errors.

### 2.3 The Five Parameter Logistic Model

The currently used 5PL function is depicted in Equation 2.3, below

\[
y = f(x; \mathbf{p}) = f(x; A, B, C, D, g) = D + \frac{A - D}{\left(1 + \left(\frac{x}{C}\right)^B\right)^g},
\]

(2.3)

- **A** = Minimum asymptote. In a bioassay where you have a standard curve, this can be thought of as the response value at a 0 standard concentration.

- **B** = Hill’s slope. The Hill’s slope refers to the steepness of the curve. This parameter controls the speed to the asymptotes. It can either be positive or negative.

- **C** = Inflection point. The inflection point is defined as the point on the curve where the curvature changes direction or signs, and must be greater than 0.
• D = Maximum asymptote. In an bioassay where you have a standard curve, this can be thought of as the response value for infinite standard concentration.

• g = Jointly with B controls the rate of approach to the D asymptote, and greater than 0.

The 4PL function is obtained by setting g=1.

After the 5PL model has been fitted to standard data, estimates for single-dilution unknown concentrations x can be obtained from unknown responses y using the inverse of equation 2.3, that is,

\[ x = f^{-1}(y; p) = f^{-1}(y; A, B, C, D, g) 
= c \left( \left( \frac{A - D}{y - D} \right) \frac{1}{g} - 1 \right) \frac{1}{B} , \]  \tag{2.4}

As we mentioned above, the extra parameter g is introduced for addressing asymmetry. Both the shape parameter B and parameter g jointly determine the speed for the response to approach the two asymptotes A and D. Setting g=1 in equation 2.2 reduces to the 4PL function in equation 2.1, and a median effect of dose(ED50) is given by equation 2.5,

\[ ED50 = C \left( 2^{1/g} - 1 \right)^{1/B} . \]  \tag{2.5}

2.4 The Effects of The Five Parameters In The Logistic Model

The 5PL function given by equation 2.3 has been used to improve the performance over the 4PL as given by equation 2.1. When we change a parameter while keeping the others fixed, the dynamic pattern of the curve change and are not the same as those of the 4PL. Thus, the parameter effect on the pattern changes of the current 5PL is displayed in the following graphs, Figures 2.2-2.6.
Figure 2.: The Effect of Parameters A in five parameters logistic model based on \( y = D + \frac{A-D}{\left(1+\frac{x}{c}\right)^B} \).

The above Figure 2 shows the behavior of parameter A while fixing B=5, C=5, D=5, and g=4. This figure also shows the difference between the five parameters and four parameters logistic curve. The IC50 value in 4PL is 5, and 3.8 in 5PL with A=0 and fixed parameters.

Figure 3.: The Effect of Parameters B in five parameters logistic model based on \( y = D + \frac{A-D}{\left(1+\frac{x}{c}\right)^B} \).

The above Figure 3 shows the behavior of parameter B while fixing A=0, C=5, D=5, and g=4. This figure also shows the difference between the five parameters and four parameters.
logistic curve. The IC50 value in 4PL is 5, and 3.8 in 5PL with B=5.5 and fixed parameters.

Figure 4.: The Effect of Parameters C in five parameters logistic model based on

\[ y = D + \frac{A-D}{\left(1+\frac{x}{c}\right)^B} \]

The above Figure 4 shows the behavior of parameter C while fixing A=0, B=5.5, D=5, and g=4. This figure also shows the difference between five parameters and four parameters logistic curve. The IC50 value in 4PL is 5, and 3.8 in 5PL with C=5 and fixed parameters.

Figure 5.: The Effect of Parameters D in five parameters logistic model based on

\[ y = D + \frac{A-D}{\left(1+\frac{x}{c}\right)^B} \]

The above Figure 5 shows the behavior of parameter D while fixing A=0, B=5.5, C=5,
D=5, and g=4. This figure also shows the difference between five parameters and four parameters logistic curve. The IC50 value in 4PL is 5, and 3.8 in 5PL with D=4 and fixed parameters.

\[
y = D + \frac{A-D}{\left(1+\left(\frac{x}{C}\right)^{B}\right)^{g}}.\]

Figure 6.: The Effect of Parameters g in five parameters logistic model based on

The above Figure 6 shows the behavior of parameter g while fixing B=5, C=5, D=5, and g=4. This figure also shows the difference between five parameters and four parameters logistic curve. The IC50 value in 4PL is 5, and 3.8 in 5PL with A=0 and fixed parameters.

2.5 Development Of A Statistical Algorithms To Estimate The Five Parameters Logistic Model

Dose response data can either be asymmetric or symmetric; therefore, we could not apply four parameters logistic model directly. We proceeded to develop a statistical algorithm to estimate the five parameters of the logistic model, 2.2. We will accomplish this in developing the algorithm that will be based on three parts of our modules. A schematic diagram of the overall proposed drug efficiency modeling in shown by Figure 7.
The overall structure of the main modules consists of three parts:

- Outlier Replicate Remover.
- Dynamic Initial Value Detection.
- Estimate the parameters of logistic model.

The subject data depends on several replicated experiments; however, there are some unexpected results that we call outliers. We use the distance algorithm between replicated experiments to detect outliers. In the parameter estimation, we have the initial value problem because we need to detect the local maximum or local minimum. This is a very common problem in the estimation procedure. We could easily avoid this problem by using all possible combinations of the initial starting points, but it is not an efficient way to estimate parameters. Thus, we proceed to apply dynamic initial values in our program for quick and accurate estimates.
The schematic diagram shown by Figure 2.7, shows how we detect the dynamic initial value.

![Dynamic Initial Value Detection Diagram]

Figure 8.: Dynamic Initial Value Detection.

At this point, we need to find the best initial points of our data and proceed to develop a non-linear statistical regression model. A good estimate of the parameters depends on having the best possible initial points. We used the following basic initial values: minimum response value for A, high response value for D, middle dose value for C, 5 for B, and 0.5 for g. Using those basic initial values, we applied them to a five parameters logistic model using non-linear regression for the first replicated data through the last replicated data. Before we used the replicated dose data, we also checked for outliers using the distance theorem. We used the mean of each of the estimated parameters of replicated data for the final initial values for the combined data set. Figure 9 shows the algorithm of the Drug Efficiency Estimate program.

- Step 1. Use basic initial value to apply the five parameters logistic model on each replicate result.

- Step 2. Save all estimated parameters for the dynamic initial value.

- Step 3. Apply the dynamic initial value to combined data set to estimate the five parameters.

- Step 4. If the minimum asymptote is less than zero, we apply the three parameters logistic model.

- Step 5. calculate IC50 and ED50 value.
2.6 Drug Efficiency Estimator Program

We made a graphic user interface (GUI) standalone program, which also included sample data. We used a non-linear regression fit, the dynamic changing initial data points for logistic models, and outlier detection for replicated experience. For outlier detection, we used the distance theorem, and five, four, and three parameters logistic regression models with the non-linear regression estimator.

Our GUI program has three parts. It consists of a reading data part, drug experiment part, saving result, scatter plot, and result curve plot. Data format must be followed by the sample data format. The sample data is in a sample folder of our program package. Figure 10 shows the main figure of Drug Efficiency Estimate Program, and Figure 11 shows the result of the Drug Efficiency Estimate Program.

The sample data which we used to get Figure 11 consisted of 12 independent drugs’ response data with three replicated experiments for each of the drugs. It has 10 dose levels.

Figure 10.: Main Figure of Drug Efficiency Estimate Program.
Using this graphical result, we can see which drug(s) need more experimentation. For example, Cell response of Drugs 2, 6, 9, 10, 11, 12 reach 0% before they have a minimum asymptote. In this case, medical scientists need to conduct more experimentation to estimate their IC50 and ED50 values.

2.7 Compare with Commercial Program

We compared our program to the GraphPad Prism[13] which is a commercial program developed by GraphPad Software, Inc. There are several functions related to statistical comparisons, column statistics, lineage regression and correlation, nonlinear regression, and clinical lab statistics. For comparison purposes, we focused on the nonlinear regression aspect. The GraphPad also includes an automatic outlier elimination and support replicates, but the program uses only one mathematical model to estimate parameters; therefore, there are large lack-of-fit errors when we apply the four parameters logistic model to asymmetrical data. Our program can estimate all the parameters for symmetric or asymmetric data,
including non zero concentration. GraphPad doesn’t support dynamic initial values, and it requires a manual input of basic initial values. The reason why manual input can pose a problem is because there are un-identified parameters problems associated with it.

2.8 Summary and Contribution

In this Chapter, we introduced the four parameters logistic model, equation 2.1 and the five parameters logistic model, equation 2.3. We showed the graphical parameters effect on the five and four parameters logistic model. The four parameters logistic model is a special case in five parameters logistic model when \( g = 1 \); however, we can only apply the four parameter logistic model for symmetrical data. Therefore, there are many restrictions in the four parameters logistic model because almost all the data is not symmetrical. Thus, the purposed five parameters logistic model can handle asymmetrical medical data. There are two types of resistance: the first one is when parameter \( B \) reaches 0 in the middle of the dosage. In this case, we can’t estimate the IC50, but we can get the ED50. The second type of resistance occurs when we have a negative estimated \( B \) parameter. In this case, we proceed to estimate the parameters using the three parameters logistic model which is a special case in the four parameters logistic model when \( B = 0 \). This program will give medical scientists a scatter data plot, a fitted regression curve, and a result file which includes estimated parameters, indications for outliers, and indications for a need to continue experimentation. Medical scientists can reduce their time to obtain significant values for IC50 or ED50 and easily view and compare the other new drug’s efficiency results. We also provide the excel result file generated by our Drug Efficiency Estimator GUI program.

1. Obtain the optimal dose of a given drug using our Drug Efficiency Estimator GUI program.

2. Dynamic initial start points for estimations.

3. Graphical and text result generated by our program.
4. It calculate the most significant IC50 and ED50.

5. Outlier detection for unexpected experimentation using the distance algorithm.

6. Medical Scientists can easily use our GUI program to view and to compare their new drug’s response.
Chapter 3

Survival Analysis of Breast Cancer Data Using the Random Forests: an Ensemble of Trees

3.1 Introduction

Tree-based methods by Bell(1999),[2] have become popular for performing survival analysis with complex data structures such as the SEER data obtained from SEER website.[8]. Within the Random Forest by Liaw(2002),[19], he applied the decision tree analysis by Smith(2004),[29] to identify the most important attributable variables that significantly contribute to estimating the survival time of a given cancer patient and proceeded to develop a statistical model to estimate the survival time.

In this chapter, we present data mining techniques to make classified groups and rank attributable variables using a weighted counting method. Data mining is the process of analyzing large quantities of data and summarizing it into useful information.

According to WHO statistical reports by D.M. Parkin (2005),[6] the incidence of breast cancer is the number one form of cancer among women. In the United States, approximately one in eight women over their lifetime have a risk of developing breast cancer. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis, AMSWebite,[30]. Breast cancer occurs due to an uncontrolled growth of cells in the breast tissues by BCF,[3] and damage to a cell’s DNA, National Breast Cancer Center,[22]. A tumor is an abnormal cell growth that can either be benign or malignant. Benign tumors are non invasive while malignant tumors are cancerous and spread to other parts of the body. Early diagnosis and treatment help to
prevent the spreading of cancer.

We are using breast cancer in the SEER data and have introduced the classification approach according to our research interest.

In this chapter, we introduce the new decision tree algorithm and weighted counting method for the final tree and rank of variables.

### 3.1.1 Introduction to Decision Tree

As simple as a decision tree may be, it is a very powerful form of multiple variable analysis used with medical data, and especially cancer research. This is an extension of multiple linear regression, and one of the data mining tools and techniques. In addition, decision tree analysis has been used in many fields including financial, engineering, sports, and presidential elections. It is produced by algorithms that identify the point for splitting a data set into two groups. A decision tree can be used for classification (categorical variables) or regression (continuous variables) applications. Rules are developed using software available in many statistics packages. Different algorithms are used to determine the "best" split at a node. We used the log rank test and the Kolmogorov-Smirnov test(K-S test) to find the best split at a node.

A decision tree consists of a root node and branches. The root decision node is the starting point of the tree, and the branches connect to their child nodes. If the child nodes do not have their own child nodes, we called it a Terminal node. The general form of a decision tree is illustrated in Figure 3.1.
We can interpret the basic form of a sample decision tree in the following figures. The above Figure 12 has three classified groups. The first group is consistent with patients in stages I, II, and III at ages less than 50. The second group is consistent with patients in stages I, II, III at ages greater than or equal to 50. The last group is consistent with patients in stage IV.

There are several advantages to using a decision tree over other classification theory tools such as neuro networking, because of its easy interpretability, simplicity, and short time analysis.

### 3.1.2 Theory Behind Decision Tree Analysis

The basic idea of a decision tree analysis is to split the given source data set into subsets by recursive portioning of the parent node into child nodes. This is based on the homogeneity of within-node instances or separation of between-node instances with respect to target variables. For each node, attributes are examined and the splitter is chosen to be the attribute such that after dividing the nodes into two child nodes, according to the value of the attribute variable, the target variable is differentiated to be the best using an algorithm.
Because of this, we need to be able to distinguish important attributes and those that contribute little to the overall decision process. This process is repeated onto each child node in a recursive manner until splitting is either non-feasible or all certain pre-specified stopping rules are satisfied. Rpart, which is one of the R packages related to the decision tree, uses complexity parameter (CP) to stop splitting.

Classification & Regression Tree is a decision tree algorithm (L. Breiman [18], 1984) that is a non-parametric probability distribution tree technique that constructs binary classification or regression trees. Splitting points on attribute variables and choosing the best variable are chosen based on Gini impurity and Gini gains following interesting interpretation. If one object is selected at random from one of the C classes, according to the probability \((p_1, p_2, p_3, \ldots, p_c)\), it is randomly assigned to a class using the same distribution. The probability of misclassification is given by,

\[
\text{Gini index } p = \sum_{j \neq k} P_{ij}P_{ik} = 1 - \sum_k P_{ik}^2 \tag{3.1}
\]

where \(k_i\) is the most frequent class in node \(i\).

If the target variable is continuous, the split criterion is used with the Least Squares Deviation (LSD) as the impurity measure. If there is no Gini gain or the preset stopping rule is satisfied, the splitting process stops. The CHAID (Chi-Squared Automatic Interaction Detection) classification technique, introduced by Kass[17] (1980) for nominal predictors and extended by Magidson [21] (1993) to ordinal predictors, is another effective approach for nominal or ordinal target variables. CHAID exhausts all possible categorical pairs of the target variable and merges each pair until there is no statistically significant difference within the pair using the Chi-square test.

Suppose there are only two outcomes "Yes" or "No" in the root node T of the target variable. Let \(p\) and \(n\) denote the number of positive records and negative records, respectively. The initial information entropy is given by,
\[ I(p,n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}. \] (3.2)

If attribute \( X \) with values \( x_1, x_2, ..., x_n \) is chosen to be the split predictor and partitions the initial node into \( \{T_1, T_2, ..., T_N\} \) and \( p_i \) and \( n_i \) denotes the number of positive records and negative records in the child node \( i \), then the expected information \( EI(X) \) and information gain \( G(X) \) are given by,

\[ EI(X) = \sum_{i=1}^{N} \frac{p_i + n_i}{p+n} I(p_i,n_i) \] (3.3)

and

\[ G(X) = I(p,n) - EI(X), \] (3.4)

### 3.1.3 Random Forest

Random forest is an ensemble learning method developed by Leo Breiman and Adele Cutler \[?\] in 2001. Random forest offers more accuracy than the single-tree method, and it gives us what variables are important in the classification. It can handle very large databases and thousands of input variables without variable deletion. The main idea is resampling and simulation. Using bootstrap sample \( D_i \) from training data set ( \( D \) ) with replacement, we construct tree \( T_i \) using \( D_i \). At each node, we choose a random subset of features, and only consider splitting on those features. Each tree gives us information for prediction or classification. With all the information gained from the trees, we build the final tree to interpret our data. We can also find the rank of variables using the weighted counting method.

The following Figure 13 shows how to build the random forest using many decision trees.
In the random forest building process, we use the following steps:

- Step 1. Select a variable
• Step 2. Divide two groups and compare them using the log-rank test or K-S test.

• Step 3. Find the best split point on the selected variable.

• Step 4. We can also(Steps 1 4).

• Step 5. Decide the best variable and the split point which is the root node.

• Step 6. Steps 1 5 Repeating steps 1-5 using left child node data.

• Step 7. Steps 1 5 Repeat steps 1-5 using right child node data.

• Step 8. Continue to do Steps 1 7 until terminal nodes are reached.

In our R package, we have three options to choose the best split point and select variable. The basic option is "Rpart" which is the most popular function in a decision tree. The second option is "LRT" which uses the log-rank test. The third option is "KS" which uses the Kolmogorov-Smirnov Test. In this chapter, we used the basic option to build the final tree and ranking variables.

3.2 About Data

In the present study, we use breast cancer data given in the surveillance, epidemiology and end results(SEER). The SEER data[8] is reliable because it is supported by the National Cancer Institute(NCI). The SEER breast cancer incidence data consists of three datasets named YR1973_2008.SEER9, YR1992_2008.sj_la_rg_ak, YR2000 _ 2008.ca_ky_lo_nj_ga, YR2005.lo_2nd_half. The SEER currently collects and publishes cancer incidences and survival data from population-based cancer registries covering approximately 28 percent of the US population as seen on the SEER Website[8]. Furthermore, the SEER database is a premier source for cancer statistics in the United States, containing information on incidence, prevalence and survival from specific geographical areas of the US population. It also contains cancer mortality information for the entire country.
Breast cancer occurs due to an uncontrolled growth of cells in the breast tissues. A tumor is an abnormal cell growth that can either be benign or malignant. Benign tumors are non-invasive while malignant tumors are cancerous and spread to other parts of the body. An analysis of the most recent data has shown that the survival rate is 85.98% after 5 years and 80% after 10 years of diagnosis. Early diagnosis and treatments help to prevent the spreading of cancer and improve the survival rate. Using random forests, we will make classified groups to determine 5-year survival rate.

The ethnicity data size is shown by Figure 14 below.

![SEER Breast Cancer by Race (1988~2006)](image)

Figure 14.: SEER Breast Cancer Diagram.

SEER 9 Breast cancer data includes three types of races. There are 496,153 Caucasians, 17,207 African Americans, and 33,434 Asians. The following Figure 15 shows the three races’ information including three types of breast cancer which are ductal, medullary, and lobular.

Ductal cancer is the most common type of breast cancer. According to the breast-cancer.org, about 80% of all breast cancers are invasive ductal cancer. In our data set, nearly 81% of cancers ductal breast cancer in Caucasians, about 83% in African American, and 87% in Asian.
The following Figure 17 shows a systematic data diagram for breast cancer in Caucasians. Young and Chunling (2010) found that the age 50 is an important factor in breast cancer. They also proved that there is a significant different among survival times for each
3.3 Data pre-processing Methods

Data pre-processing by Han(2000),[14] is a significant issue in data mining studies. Data preprocessing includes data cleaning, data transformation and data reduction. It is also the process of discovering new patterns from large data sets by Chakrabarti(2009)[31]. Classification is a data mining technique based on machine learning which is used to classify each item in a set of data into a set of predefined classes or groups by Han(2000),[14]. Classification methods in data mining make use of mathematical and statistical techniques such as decision trees, linear programming, neural network and support vector machines. In the present study, we used the data YR1073_2008.SEER9 for breast cancer data which consists of 630,218 records with 124 variables. We will use the random forest technique that will be discussed in the next section to identify and rank the most significant attributable variables for survival time.

Figure 17.: SEER Breast Cancer Diagram : Caucasians.
3.4 Pre-Processing

Data pre-processing was applied to SEER data to prepare the raw data. Pre-processing is an important step that is used to transform raw data into a format that makes it possible to apply data mining techniques and to improve the quality of data[14]. It should be noted from related work, that attribute selection plays an important role in identifying parameters that are important and significant for proper breast cancer diagnosis. Data pre-processing is followed by addressing the missing value and identifying or removing outliers. For example, the tumor size variable contains values greater than 900 mm, which is abnormal and thus, incorrect because of data entry errors, and; therefore, are removed. Another example is that an extension of disease is an important factor in breast cancer survivability. However,32.25% of our data was missing and unknown, therefore those records were also removed. The following diagram shows steps to how we removed variables.

![Diagram showing step-by-step procedure to remove non-significant variables]

Figure 18.: Step-by-Step Procedure to Remove Non Significant Variables.

Feature selection is also implemented with logistic regression backward selection conducted by using p-value > 0.1 as the exit criterion and p-value < 0.05 as the entry criterion. After using these in data preprocessing, the final 6 discrete input variables are shown in...
Table 1. The final 7 numeric input variables are shown in Table 2. Predicting breast cancer survivability is a binary classification problem, "survived" or "not survived". If survival time recorded is equal to or more than 5 years, survival state indicates 0. If survival time recorded is less than 5 years and the cause of death is breast cancer, survival status indicates 1, which means they did not survive. The dependent variable is survival status. Distribution of the dependent variable is shown in Table 1. After data pre-processing there are 13 risk factors, and 1 dependent variable named survival status. The total is 82,128 records and 13 variables. The percentage of "survived after 5 years" is 85.98% while it is reported in NCI’s official data that in the United States the percentage of "survival after 5 years" was 88% in the year 2006.

Table 1: Description of Discrete Variables.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Field Description</th>
<th>Number of Distinct Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR_ STAT</td>
<td>Marital Status at Diagnosis</td>
<td>6</td>
</tr>
<tr>
<td>GRADE</td>
<td>Grade</td>
<td>5</td>
</tr>
<tr>
<td>D_ AJCC_S</td>
<td>AJCC Stage Group</td>
<td>11</td>
</tr>
<tr>
<td>RAD_SURG</td>
<td>Radiation Sequence with Surgery</td>
<td>7</td>
</tr>
<tr>
<td>RAC_RECY</td>
<td>Race</td>
<td>6</td>
</tr>
<tr>
<td>FIRSTPRM</td>
<td>First Malignant Primary Indicator</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Basic Summary of Discrete Variables.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Field Description</th>
<th>Mean</th>
<th>S.D.</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE_DX</td>
<td>Age at diagnosis</td>
<td>60.41</td>
<td>13.64</td>
<td>12-99</td>
</tr>
<tr>
<td>EOD10_PN</td>
<td>Positive lymph nodes examined</td>
<td>1.15</td>
<td>3.13</td>
<td>0-84</td>
</tr>
<tr>
<td>EOD10_NE</td>
<td>Number of lymph node examined</td>
<td>3.26</td>
<td>3.14</td>
<td>0-39</td>
</tr>
<tr>
<td>CS_SIZE</td>
<td>CS Tumor size</td>
<td>20.75</td>
<td>20.98</td>
<td>0-922</td>
</tr>
<tr>
<td>CS_EXT</td>
<td>CS Extension</td>
<td>0.89</td>
<td>7.67</td>
<td>0-70</td>
</tr>
<tr>
<td>CS_NODE</td>
<td>CS Lymph Nodes</td>
<td>0.26</td>
<td>3.56</td>
<td>0-50</td>
</tr>
<tr>
<td>NUMPRIMS</td>
<td>Number of Primaries</td>
<td>1.25</td>
<td>0.52</td>
<td>1-6</td>
</tr>
</tbody>
</table>

Figure 19 shows us the process of getting our final model for each of the groups.

After we have cleaned the data set, the first step is to select 80% of the sample from
cleaned data. In this sampling, we used seed number 1234. Using 80% of the data, we performed a bootstrapping and the decision tree method on our data. Our response variable is a binary factor; therefore, we used the class method using Rpart which is a well known package for a decision tree in R. The result of Rpart doesn’t include layer information, so we need to use parsing which is a data mining technique in computer science. We combined the C++ and R to do result parsing for selecting and ranking risk factors. A frequency method is used to identify and select each layer’s variable(s) and find the best point of each node. Once we obtained the final tree’s classification points, we classified our data using end node information. Table 3 shows us the rank of our risk factors from the first simulation.

The grade, number of primaries, race, and first malignant primary indicator are selected less than 100 times during 1,000 simulations. Compared to other variables, those are not important risk factors, so we do another 1,000 simulations using high rank risk factors which are selected over 100 times as shown in Table 4. The order of rank is the same as a previous simulation shown in Table 3. With this result, we confirmed that the order of important risk factors is significant.
Table 3: Rank for Risk Factors.

<table>
<thead>
<tr>
<th>Number</th>
<th>Variable Field Description</th>
<th>Freq.Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AGE_DX Age at diagnosis</td>
<td>549</td>
</tr>
<tr>
<td>2</td>
<td>CS_SIZE CS Tumor size</td>
<td>543</td>
</tr>
<tr>
<td>3</td>
<td>D_AJCC_S AJCC Stage Group</td>
<td>481</td>
</tr>
<tr>
<td>4</td>
<td>CS_NODE CS Lymph Nodes</td>
<td>277</td>
</tr>
<tr>
<td>5</td>
<td>EOD10_NE Negative lymph nodes examined</td>
<td>220</td>
</tr>
<tr>
<td>6</td>
<td>EOD10_PN Positive lymph nodes examined</td>
<td>220</td>
</tr>
<tr>
<td>7</td>
<td>MAR_STAT Marital Status at diagnosis</td>
<td>196</td>
</tr>
<tr>
<td>8</td>
<td>RAD_SURG Radiation sequence with surgery</td>
<td>129</td>
</tr>
<tr>
<td>9</td>
<td>GRADE Grade</td>
<td>96</td>
</tr>
<tr>
<td>10</td>
<td>NUMPRIMS Number of primaries</td>
<td>33</td>
</tr>
<tr>
<td>11</td>
<td>RAC_RECY Race</td>
<td>22</td>
</tr>
<tr>
<td>12</td>
<td>FIRSTPRM First malignant primary indicator</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 4: Summary of Rank for Risk Factors.

<table>
<thead>
<tr>
<th>Number</th>
<th>Variable Field Description</th>
<th>Average Freq.Sum</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AGE_DX Age at diagnosis</td>
<td>599</td>
<td>23.8</td>
</tr>
<tr>
<td>2</td>
<td>CS_SIZE CS Tumor size</td>
<td>557</td>
<td>19.2</td>
</tr>
<tr>
<td>3</td>
<td>D_AJCC_S AJCC Stage Group</td>
<td>527</td>
<td>19.1</td>
</tr>
<tr>
<td>4</td>
<td>CS_NODE CS Lymph Nodes</td>
<td>288</td>
<td>15.3</td>
</tr>
<tr>
<td>5</td>
<td>EOD10_NE Negative lymph nodes examined</td>
<td>272</td>
<td>15.2</td>
</tr>
<tr>
<td>6</td>
<td>EOD10_PN Positive lymph nodes examined</td>
<td>196</td>
<td>14.7</td>
</tr>
<tr>
<td>7</td>
<td>MAR_STAT Marital status at diagnosis</td>
<td>140</td>
<td>13.9</td>
</tr>
<tr>
<td>8</td>
<td>RAD_SURG Radiation sequence with surgery</td>
<td>130</td>
<td>13.2</td>
</tr>
</tbody>
</table>
Figure 20.: Final Tree.

Figure 21.: Survival Curves using Final Classified Groups.
We can clearly see different survival curves for the classified groups. G1 represents ages that are less than 33, G2 shows ages that are less than 65 and greater than or equal to 33 and marriage state is two. G3 shows age less than 65 and greater than or equal to 33 and marriage state is not two. G4 shows stage 1 and 2. G5 shows stage 3 and 4. The P-value in log rank test is 0.03.

3.5 Validation of Procedure

For validation, we used a cross-validation method. We have data for two groups which are the 80% data group, which we applied to our new approach, and the 20% data group. Each group has 5 sub-groups. Logrank test is a statistical hypothesis test, where the hypothesis is that the 2 groups have the same survival distribution. In particular, we divide the time into \( m \) intervals. Let \( n_{kj} \) be the number of individuals that are alive in group \( k \) at the beginning of time interval \( j \). Let \( d_{kj} \) be the number of events occurring in group \( k \) during interval \( j \). Also, \( n_j \) and \( d_j \) are defined as Eq(3.5) and Eq(3.6), respectively.

\[
    n_j = \sum_{k=1}^{n} n_{kj}, \quad \text{(3.5)}
\]

and

\[
    d_j = \sum_{k=1}^{n} d_{kj}. \quad \text{(3.6)}
\]

The test statistic is calculated by the chi-square test, that is,

\[
    \chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}, \quad \text{(3.7)}
\]

where \( O_k \) represents the number of observed deaths in group \( k \), i.e.,

\[
    O_k = \sum_{j=1}^{m} d_{kj}, \quad \text{(3.8)}
\]
and $E_k$ is the expected number of deaths in group $k$, i.e.,

$$E_k = \sum_{j=1}^{n} \frac{n_k d_j}{n_j}.$$  \hspace{1cm} (3.9)

We performed 100 simulations and log-rank tests. The result of the log rank test for the 80% data and 20% data using random forest simulation is shown in Table 5.

Table 5: Log Rank Test.

<table>
<thead>
<tr>
<th>Classified Group</th>
<th>Average P-value</th>
<th>Average S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1: $Age_{dx} \leq 33$</td>
<td>0.083</td>
<td>0.01</td>
</tr>
<tr>
<td>Group2: $33 \leq Age_{dx} \leq 65, MAR_STAT=2$</td>
<td>0.077</td>
<td>0.008</td>
</tr>
<tr>
<td>Group3: $33 \leq Age_{dx} \leq 65, MAR_STAT \neq 2$</td>
<td>0.071</td>
<td>0.009</td>
</tr>
<tr>
<td>Group4: $D_AJCC_S == 1or2$</td>
<td>0.061</td>
<td>0.012</td>
</tr>
<tr>
<td>Group5: $D_AJCC_S == 3or4$</td>
<td>0.082</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Following the result of the simulation, the average P-value of each group is greater than 0.05 and the average standard deviation is less than or equal to 0.01. There is no difference between the two data sets which we used for our result and the rest of data.

### 3.6 Summary and Contribution

In this chapter, we proposed a new methodology using a randomized decision tree utilizing bootstrapping method for partitioning tree-based models to fit the SEER Breast Cancer Dataset. Our proposal optimizes the way to rank our risk factors and to find the best classified points and risk factors for each node. We proposed a new pruning rule which is based on minimizing the tree’s overall mean squared error, where labeling rule is based on identifying sub-groups of specific survival models. Using the Breast Cancer data, we validated our method using simulation and a cross validation method.
• Developed a new decision tree algorithm based on log-rank test and K-S test.

• Developed a new algorithm to rank risk factors using count and weighted count method.

• Proposed a method to validate our new method using cross validation and simulation.

• Developed generalized R functions which researchers can use and apply to all types of data sets.
Chapter 4

4.1 Introduction

The four parameter Johnson SB probability distribution is important in parametrically studying a variety of health, environmental, and engineering problems. Scientists today are experiencing difficulties in obtaining estimates for the four parameters of the subject probability distribution.

Many methods have been developed to estimate the parameters of the Johnson probability distribution, such as the four percentile method, Wheeler, Slifker and Shapiro [28], 1980. The Knoebel-Burkhart method (Knoebel and Burkhart [27], 1991) is another method to obtain approximate estimates.

The aim of the present study is to review the existing methods to estimate the four parameters of the subject probability density function and to propose a fast algorithm that produces better approximations for these parameters. The usefulness of this procedure is illustrated using numerical simulation and real cancer data.

In the present study, we developed functions in the R package to generate random numbers and estimates for the four parameters, draw graphics, and analyze cancer data using the Johnson SB probability distribution. We used normal transformations to generate random numbers and the Newton-Raphson method to estimate the four parameters. In addition, we incorporated general integration and a prediction procedure to draw useful graphs using R.
4.2 Numerical Methods

In 1949, Johnson\[16\] proposed a system of three probability distributions based on transformations of a standard normal probability distribution. The general form of the Johnson system of distributions is given by

\[
Z = \gamma + \delta f(x; \xi, \delta),
\]

with \( \delta > 0, -\infty < \gamma < \infty, \delta > 0, -\infty < \xi < -\infty. \) (4.2)

Where \( Z \) is a standard normal variable and the random variable \( x \) followed the normal probability density function \( f(x) \). \( Z \) must be specified, and it depends on a set of chosen parameters. Bounded by \((\xi, \xi + \lambda)\), The parameters \( \gamma \) and \( \delta \) are the shape parameters, \( \xi \) is the location parameter, and \( \lambda \) is the scale parameter. The values of these parameters determine the shape and behavior of the probability distribution. Johnson suggested the following functions, denoted by \( S_L, S_B, \) and \( S_U \), respectively,

\[
f(x; \xi, \gamma) = \ln \left( \frac{x - \xi}{\gamma} \right), \quad x > \xi, \quad \text{(4.3)}
\]

\[
f(x; \xi, \gamma) = \ln \left( \frac{x - \xi}{\gamma + \xi - x} \right), \quad \xi < x < \xi + \gamma, \quad \text{(4.4)}
\]

and

\[
f(x; \xi, \gamma) = \sinh^{-1} \left( \frac{x - \xi}{\gamma} \right), \quad -\infty < x < \infty, \quad \text{(4.5)}
\]

We are interested in using the function given in equation (4.4), with the transformation given by (4.1) to obtain,
\[ Z = \gamma + \delta \ln \left( \frac{x - \xi}{\gamma + \xi - x} \right), \quad \xi < x < \xi + \lambda, \quad (4.6) \]

Since \( Z \) is a normally distributed random variable, the variable \( x \) is also identified as log-normally distributed. The probability density function of \( x \) under the transformation, (4.6), is given by

\[ f(x) = \frac{\delta}{\lambda \sqrt{2\pi y(1-y)}} \exp \left[ -\frac{1}{2} \left\{ \gamma + \xi \left( \frac{y}{1-y} \right)^2 \right\} \right], \quad y = \frac{x - \xi}{\lambda}. \quad (4.7) \]

Depending on the parameter values, the Johnson SB probability distribution has one or more modes. Figure 22, displays the form of the distribution with the location set at 0 and 100 when the parameters \((\gamma, \delta)\) have the values \((0, 0.5), (0, 0.75), (0.25, 0.5), (0.5, 0.5), (1, 2),\) and \((1, 1)\).

![Figure 22: Johnson’s SB distribution with \((\gamma, \delta)\) combinations first row-left(0,0.5) ,first row-middle(0,0.75), first row- right (0.25,0.5), second row-left(0.5,0.5),second row-middle (1,2),second row-right (1,1).](image)

In the JohnsonUSF package, we used the Newton-Raphson method to estimate the pa-
rameters $\gamma$ and $\delta$. Using equation 4.7, the first and second partial derivatives with respect to $\gamma$ and $\delta$ are given by

\[
\frac{\partial L}{\partial \gamma} = -\sum_x \left( \gamma + \delta \times \log \left( \frac{y}{1-y} \right) \right),
\]

(4.8)

\[
\frac{\partial L}{\partial \delta} = \frac{n}{\delta} - \sum_x \left( \gamma + \delta \times \log \left( \frac{y}{1-y} \right) \times \log \left( \frac{y}{1-y} \right) \right),
\]

(4.9)

\[
\frac{\partial^2 L}{\partial \gamma^2} = -n,
\]

(4.10)

\[
\frac{\partial^2 L}{\partial \gamma \partial \delta} = -\sum_x \log \left( \frac{y}{1-y} \right),
\]

(4.11)

and

\[
\frac{\partial^2 L}{\partial \delta^2} = -\frac{n}{\delta^2} - \sum_x \left( \log \left( \frac{y}{1-y} \right) \right)^2,
\]

(4.12)

In the JohnsonUSF package, the numerical procedure to estimate the four parameters of the Johnson SB probability distribution(6) is displayed in the following steps.

Step 1 : Estimate $\xi$ and $\lambda$ using a quantile method.

Step 2 : Estimate $\gamma$ and $\delta$ using a algorithm flow chart : Figure 23.

In the implementation of the algorithm, $(\beta_1, \beta_2)$ represents the values of the vector $(\gamma, \delta)$ and is recursively updated in $(\beta_1^*, \beta_2^*)$ using the following equation:

\[
\begin{pmatrix}
\beta_1^* \\
\beta_2^*
\end{pmatrix} = \begin{pmatrix}
\beta_1 \\
\beta_2
\end{pmatrix} - H^{-1} \times S,
\]

(4.13)

where the Hessian Matrix H is given by

42
Figure 23.: Algorithm to estimate the four parameters of the Johnson SB Probability Distribution.
\[ H = \begin{pmatrix} \frac{\partial^2 L}{\partial \gamma^2} & \frac{\partial^2 L}{\partial \gamma \partial \delta} \\ \frac{\partial^2 L}{\partial \gamma \partial \delta} & \frac{\partial^2 L}{\partial \delta^2} \end{pmatrix} \quad (4.14) \]

and the score vector \( S \) is given by

\[ S = \begin{pmatrix} \frac{\partial L}{\partial \gamma} \\ \frac{\partial L}{\partial \delta} \end{pmatrix} \quad (4.15) \]

### 4.3 How to Use the JohnsonUSF Package

The JohnsonUSF package can be installed via the Install package(s) from the local zip files option of the Packages menu (view Figure 24)

![Figure 24. Using local JohnsonUSF.zip file to install on your computer.](image)

There are four parts in the JohnsonUSF package which generate random numbers, find probability, critical values, and estimated parameters using several statistical methods.
Table 6: Functions in JohnsonUSF Package.

<table>
<thead>
<tr>
<th>Description</th>
<th>Name of function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generate Random Number</td>
<td>rjohnsonsb(N,γ,δ,λ,ξ,method)</td>
</tr>
<tr>
<td>Calculate Probability</td>
<td>qjohnsonsb(q,tail)</td>
</tr>
<tr>
<td>Critical Value</td>
<td>qjohnsonsb(p,tail)</td>
</tr>
<tr>
<td>Estimate Parameters</td>
<td>est.johnsonsb(data,pl,method)</td>
</tr>
</tbody>
</table>

4.4 Validation of Estimated Parameters

We use the Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and Chi-Squared goodness of fit tests for our random sample generated by the JohnsonUSF package. The first statistical approach is based on the empirical cumulative distribution (ECD). The statistic used in the K-S test relies on the largest vertical difference between the ECD and theoretical function.

The second test is a modification of the K-S test and is not distribution-free since the critical value of its statistic uses a specific distribution. The third test shows how the observed value of a given "treatment" is significantly different from the observed value. For example, using values $\delta = 3$, $\gamma = 4$, $\xi = 100$, and $\lambda = 0$, ten thousand data sets of sample size 1000 were generated. We proceed to estimate the parameter vector $(\delta, \gamma)$ using the JohnsonUSF package. The mean and standard deviation for each of the parameters are given below in Table 7.

Table 7: Basic Statistics for the Parameters $\delta$ and $\gamma$ for Ten Thousand Johnson SB Probability Distributed Random Samples.

<table>
<thead>
<tr>
<th></th>
<th>$\delta = 3$</th>
<th>$\gamma = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.004595</td>
<td>4.005728</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.0740135</td>
<td>0.0898567</td>
</tr>
</tbody>
</table>
The following table shows result summaries of the K-S test for a sample with \((\delta, \gamma) = (3, 4)\)

Table 8: Kolmogorov-Smirnov, Anderson-Darling, and Chi-Square goodness of fit test results.

<table>
<thead>
<tr>
<th>Test</th>
<th>Critical Value</th>
<th>Alpha</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>0.01542</td>
<td>0.05</td>
<td>Fail to Reject H0</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>2.5018</td>
<td>0.05</td>
<td>Fail to Reject H0</td>
</tr>
<tr>
<td>Chi-Squared</td>
<td>21.0216</td>
<td>0.05</td>
<td>Fail to Reject H0</td>
</tr>
</tbody>
</table>

and Figure 25, below displays the histogram for the same random sample:

![Histogram of data](image_url)

Figure 25.: Random Sample from JohnsonSB Probability Distribution \(\delta = 3, \gamma = 4, \xi = 100, \lambda = 0\).
4.5 Apply to Real Lung Cancer Data

We used the JohnsonUSF package to estimate the parameters of the Johnson SB distribution using real lung cancer data. The sample size for this data set was 7,754 corresponding to lung cancer survival times. In addition, we employed the Easyfit program in order to perform comparisons with respect to the results obtained using the JohnsonUSF package. However, Easyfit will not give us any standard error information on each estimation nor can it carry out a simulation study. The results are summarized in the following table:

Table 9: Estimate Comparison Using JohnsonUSF Package and Other Software Program.

<table>
<thead>
<tr>
<th></th>
<th>δ</th>
<th>γ</th>
<th>ξ</th>
<th>λ</th>
</tr>
</thead>
<tbody>
<tr>
<td>EasyFit Result</td>
<td>-0.133(NA)</td>
<td>0.73(NA)</td>
<td>284(NA)</td>
<td>-6.8(NA)</td>
</tr>
<tr>
<td>JohnsonUSF</td>
<td>-0.11(0.01)</td>
<td>0.72(0.031)</td>
<td>284(2)</td>
<td>-6.8(0.8)</td>
</tr>
</tbody>
</table>

The corresponding histogram and the fitted curve for the estimates found by the proposed JohnsonUSF package is shown in Figure 26.

4.6 Conclusion

We developed an R package for a four parameters JohnsonSB probability distribution. The JohnsonUSF package includes generating random numbers, calculating probability and inverse probability, and estimating four parameters for the JohnsonSB probability distribution. In a given population, we proceed to easily work with a four parameters JohnsonSB. We progress to generate a sequence of random samples, and this package utilizes a loop type of processor that can simultaneously generate random samples and estimates of categorical behavior of the distribution of each sample. This is something that the existing method lacks and it requires extra work because each random sample information has to be manually obtained. Statistician and student can use our package to easily develop research. Our package will be kept up to date and developed using the following list.
Figure 26.: Data Fitting Using Real Lung Cancer Data.
• Apply MCMC to develop a random number generator and compare with previous study.

• Apply the E-M algorithm to obtain the parameter estimation, including censoring information.

• Develop a Bayesian model and apply the E-M algorithm to get parameters estimation.

4.7 Contribution

• Developed several functions such as probability, estimation, and graph related to the JohnsonSB Probability Distribution.

• Applied several mixed algorithms to develop different functions.

• Developed R Packages using several functions.
Chapter 5

Estimating Survival Time of Baseline Lung Cancer Using Statistical Modeling

5.1 Introduction

The purpose of the present study is to develop statistical models for predicting the survival time and probability using data for lung cancer patients - USF Technical Report #100, 2010[31]. In this study, we used the number of cigarettes per day, duration of smoking, and the age at diagnosis of lung cancer to identify the significant attributable variables including their interactions, the survival time, and probability density functions for classified groups.

Cancer is the second leading cause of death and was among one of the first diseases causally linked to smoking. In the United States, cigarette smoking causes approximately 438,000 deaths each year, including deaths related to cancer, cardiovascular disease, respiratory disease and infant mortality. Compared to nonsmokers, men who smoke are about 23 times more likely to develop lung cancer and women who smoke are about 13 times more likely. Smoking is the major determinant of lung cancer and accounts for 90% of all cases. The relationship between smoking and lung cancer is well documented (Doll and Peto[9], 1978; Doll et al.[26], 2004;IARC[11],2004), along with the benefits of smoking cessation(US DHHS[23], 1990; Peto et al.[24], 2000). Many studies have focused on two hypotheses: whether the combined effect of asbestos and smoking is addictive (each factor acts independently) or multiplicative (the effect of asbestos exposure on lung cancer risk is proportional to the effect of smoking). In our research, the purpose of this study is to identify the significant attributable variables as well as all possible interactions among those variables. We also aim to construct a statistical model to predict the survival time as
a function of those attributable variables and interactions so that we will be able to predict how long a specific lung cancer patient will survive.

5.2 Lung Cancer Data with Smoking Information

Our intent is to analyze a data set consisting of 163,386 lung cancer patients with information on survival time, sex, race, smoking indicate, number of cigarettes per day, duration, and age at diagnosis. The five races indicated are Caucasian, African American, Hispanic, Asian, and other. Sufficient data on race is not included in our data set. Although our model is built almost entirely from data on Caucasians, we assume that racial differences are negligible. We also have former smoker and current smoker indicates. There is one more assumption that one’s smoking history is continuous, e.g. no former smoker stat again, and there are no 'gaps' in a person’s smoking history.

The following Figure 27 Data Diagram of Lung Cancer Patient Information.

![Figure 27: Data Diagram of Lung Cancer's Patients Information.](image)

Our minimum survival time in months is 55 and the maximum survival time is 266; the variance is 47.52; mean of average cigarettes per day is 25.04; minimum age at diagnosis
is 32 years old while the average is 60.1 years old.

5.3 Classified Groups

We want to find out whether the true survival curves are different from current smoker and former smoker including sex information. Compute the Mantel-Haenszel (1951) statistic for those classifiers to test those groups. There is strong evidence to suggest that survival time for all our groups are different, and its p-value is around 0.0001-0.0002.

We have the following four groups.

- Group 1 : Caucasian, Current Smoker, and Male.
- Group 2 : Caucasian, Current Smoker, and Female.
- Group 3 : Caucasian, Former Smoker, and Male.
- Group 4 : Caucasian, Former Smoker, and Female.

5.4 Correlation Matrix

We are addressing more than one independent variable. The collection of all zero-order correlation coefficients can be represented most compactly in correlation matrix form. In Group I, Time and CPD, DUR and AGEATIN demonstrate a positive correlation, and Time and Dur, Time and AGEATIN, DUR and CPD, CPD and AGEATIN demonstrate a negative correlation. The following tables give correlation matrices for group I.

Table 10: Correlation Matrix on Caucasian, Current Smoker, and Male (Group I).

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>CPD</th>
<th>DUR</th>
<th>AGEATIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1.00</td>
<td>0.63</td>
<td>-0.61</td>
<td>-0.88</td>
</tr>
<tr>
<td>CPD</td>
<td>0.63</td>
<td>1.00</td>
<td>-0.41</td>
<td>-0.61</td>
</tr>
<tr>
<td>DUR</td>
<td>-0.61</td>
<td>-0.41</td>
<td>1.00</td>
<td>0.82</td>
</tr>
<tr>
<td>AGEATIN</td>
<td>-0.88</td>
<td>-0.61</td>
<td>0.82</td>
<td>1.00</td>
</tr>
</tbody>
</table>
In Group II, there is zero correlation between Time and CPD, and there are two positive correlations on CPD and DUR, AGEATIN and DUR. There are three negative correlations on Time and DUR, Time and AGEATIN, CPD and AGEATIN.

Table 11: Correlation Matrix on Caucasian, Current Smoker, and Female (Group II).

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>CPD</th>
<th>DUR</th>
<th>AGEATIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1.00</td>
<td>0.00</td>
<td>-0.73</td>
<td>-0.84</td>
</tr>
<tr>
<td>CPD</td>
<td>0.00</td>
<td>1.00</td>
<td>0.47</td>
<td>-0.03</td>
</tr>
<tr>
<td>DUR</td>
<td>-0.73</td>
<td>0.47</td>
<td>1.00</td>
<td>0.79</td>
</tr>
<tr>
<td>AGEATIN</td>
<td>-0.84</td>
<td>-0.03</td>
<td>0.79</td>
<td>1.00</td>
</tr>
</tbody>
</table>

In Group III, there are two positive correlations on Time and CPD, DUR and CPD. There are three negative correlations on Time and DUR, Time and AGEATIN, CPD and AGEATIN.

Table 12: Correlation Matrix White, Former Smoker, and Male (III).

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>CPD</th>
<th>DUR</th>
<th>AGEATIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1.00</td>
<td>0.08</td>
<td>-0.60</td>
<td>-0.65</td>
</tr>
<tr>
<td>CPD</td>
<td>0.08</td>
<td>1.00</td>
<td>0.13</td>
<td>-0.22</td>
</tr>
<tr>
<td>DUR</td>
<td>-0.60</td>
<td>0.13</td>
<td>1.00</td>
<td>0.83</td>
</tr>
<tr>
<td>AGEATIN</td>
<td>-0.65</td>
<td>-0.22</td>
<td>0.83</td>
<td>1.00</td>
</tr>
</tbody>
</table>
In Group IV, there is one positive correlation on CPD and DUR, CPD and AGEATIN. There are three negative correlations on Time and CPD, Time and DUR, Time and AGEATIN.

Table 13: Correlation Matrix on Caucasian, Former Smoker, and Female (IV).

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>CPD</th>
<th>DUR</th>
<th>AGEATIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1.00</td>
<td>-0.42</td>
<td>-0.66</td>
<td>-0.74</td>
</tr>
<tr>
<td>CPD</td>
<td>-0.42</td>
<td>1.00</td>
<td>0.68</td>
<td>0.29</td>
</tr>
<tr>
<td>DUR</td>
<td>-0.66</td>
<td>0.68</td>
<td>1.00</td>
<td>0.82</td>
</tr>
<tr>
<td>AGEATIN</td>
<td>-0.74</td>
<td>0.29</td>
<td>0.82</td>
<td>1.00</td>
</tr>
</tbody>
</table>

The following figure shows the matrix of scatter plots for all groups.

Figure 28.: Matrix of Scatter Plots for All Groups.
5.5 Survival Analysis and Statistical Modeling

The general purpose of multiple regressions (Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. The general form of the multiple regression model is given by equation 1, where $y$ is the response variable which is survival time to be estimated, and $x_i(t)$ and $x_i(t)$ are the $m$ different contributing entities $i=1,2,3, m$:

$$y = \beta_0 + \sum_{i=1}^{m} \beta_i X_i(t) + \epsilon$$  \hspace{1cm} (5.1)

In the proposed model, the descriptions of the contributing variables are provided in Table 10, and these variables are denoted in equation 5.2 below. In equation 5.2, coefficients $\beta_i's$ are the contributing variables, given in Table 9, and $\epsilon$ is the random error. The estimates of the $\beta_i's$ are the key factors used to identify the significantly contributing variables.

$$\sum_{i=1}^{m} \beta_i X_i(t) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$  \hspace{1cm} (5.2)

Table 14: Variables Description .

<table>
<thead>
<tr>
<th>Y</th>
<th>Survival Time(Month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>Number of Cigarettes Per Day</td>
</tr>
<tr>
<td>$x_2$</td>
<td>Duration(year)</td>
</tr>
<tr>
<td>$x_3$</td>
<td>Age at diagnosis</td>
</tr>
</tbody>
</table>

The maximum model is defined to be the largest model (having the most predictor variables) considered at any point in the process of model selection. All other possible models can be created by deleting predictor variables from the maximum model. Created by deleting predictors from the maximum model, our model is a restriction of the maximum model.
5.5.1 Multiple Regression Model for Group I

From Figure 29, we can see that there is a positive linear relationship between time and age and duration. Although monotonic, it may well be linear. We also have quadric behavior between cigarettes per day and duration. Using this result, we can get a maximum model for Group I given by,

\[ Y_1 = \beta_0 + \beta_1 x_1^2 + \beta_2 x_1 + \beta_3 x_2^2 + \beta_4 x_2 + \beta_5 x_3 + \beta_6 x_3^2 + \sum_{i=7}^{n} \text{int}(x_1^2, x_1, x_2^2, x_2, x_3^2, x_3) + \epsilon \quad (5.3) \]

Using backward selection, we have the final model in 5.4
\[ Y_{g1} = -4202.01 + 527.82x_1 - 159.37x_2 + 84.43x_3 - 16.92x_1^2 + 3.82x_2^2 + \\
7.13x_1 : x_2 - 8.15x_1 : x_3 + 1.96x_2 : x_3 + 0.11x_1^3 - 0.03x_2 : x_1^2 + \\
0.21x_3 : x_1^2 - 0.42x_1 : x_2^2 + 0.01x_2^3 - 0.06x_3 : x_2^2 + 0.01x_1^2 : x_2^2 + \\
0.0018x_1^3 : x_3 \] (5.4)

With multiple regression comes an overall test of significance, and a multiple \( R^2 \) which is actually the value of \( r^2 \) for the measured survival time vs. the predicted survival time. Our multiple R-squared is 0.9971, and adjusted R-squared is 0.9953 for our final Group 1 model which is Caucasian, current smoker, and male.

The following Table 15 shows us part of the residual validation.

<table>
<thead>
<tr>
<th>Actual Survival Time</th>
<th>Estimated Survival Time</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>884</td>
<td>883.44</td>
</tr>
<tr>
<td>2</td>
<td>895</td>
<td>884.77</td>
</tr>
<tr>
<td>3</td>
<td>839</td>
<td>840.56</td>
</tr>
<tr>
<td>4</td>
<td>864</td>
<td>864.07</td>
</tr>
<tr>
<td>5</td>
<td>929</td>
<td>929.13</td>
</tr>
<tr>
<td>6</td>
<td>849</td>
<td>850.72</td>
</tr>
<tr>
<td>7</td>
<td>802</td>
<td>802.46</td>
</tr>
<tr>
<td>8</td>
<td>907</td>
<td>905.73</td>
</tr>
<tr>
<td>9</td>
<td>815</td>
<td>814.31</td>
</tr>
<tr>
<td>10</td>
<td>930</td>
<td>929.55</td>
</tr>
</tbody>
</table>
5.5.2 Multiple Regression Model for Group II

From Figure 30, we can see that there is a positive linear relationship between time and age and duration. Although monotonic, it may well be linear. We also have quadric behavior between cigarettes per day and duration. Using this result, we can get a maximum model for Group II given by,

\[ Y_{g2} = \beta_0 + \beta_1 x_1^2 + \beta_2 x_1 + \beta_3 x_2^2 + \beta_4 x_2 + \beta_5 x_3 + \beta_6 x_3^2 + \sum_{i=7}^{n} int(x_1^2, x_1, x_2^2, x_2, x_3^2, x_3) + \epsilon \]  

(5.5)

Using backward selection, we have the final model in 5.6

Figure 30.: Matrix of Scatter Plots for Groups2.
\[ y_{g2}^* = -9779.15 + 1898.28x_1 + 772.32x_2 + 187.12x_3 - 114.36x_1^2 - 6.06x_2^2 - 142.67x_1 : x_2 - 33.16x_1 : x_3 - 15.68x_2 : x_3 + \\
2.2x_1^3 + 8.42x_2 : x_1^2 + 1.98x_3 : x_1^2 + 1.1x_1 : x_2^2 - 0.08x_2^3 + 0.23x_3 : x_2^2 - 0.07x_1^2 : x_2^2 + 2.89 * x_1 : x_2 : x_3 \] (5.6)

With multiple regression comes an overall test of significance, and a multiple \( R^2 \) which is actually the value of \( r^2 \) for the measured survival time vs. the predicted survival time. Our multiple R-squared is 0.9946, and adjusted R-squared is 0.9906 for our final Group 2 model which is Caucasian, current smoker, and female.

The following Table 16 shows us part of the residual validation.

Table 16: Residual Validation : Group II (Mean residual : -0.023 and standard error : 0.94).

<table>
<thead>
<tr>
<th>Actual Survival Time</th>
<th>Estimated Survival Time</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 787</td>
<td>786.26</td>
<td>-0.74</td>
</tr>
<tr>
<td>2 902</td>
<td>902.68</td>
<td>0.68</td>
</tr>
<tr>
<td>3 946</td>
<td>945.04</td>
<td>-0.96</td>
</tr>
<tr>
<td>4 936</td>
<td>934.63</td>
<td>-1.37</td>
</tr>
<tr>
<td>5 782</td>
<td>781.50</td>
<td>-0.50</td>
</tr>
<tr>
<td>6 925</td>
<td>924.38</td>
<td>-0.62</td>
</tr>
<tr>
<td>7 807</td>
<td>807.94</td>
<td>0.94</td>
</tr>
<tr>
<td>8 841</td>
<td>841.23</td>
<td>0.23</td>
</tr>
<tr>
<td>9 905</td>
<td>905.67</td>
<td>0.67</td>
</tr>
<tr>
<td>10 936</td>
<td>937.43</td>
<td>1.43</td>
</tr>
</tbody>
</table>
5.5.3 Multiple Regression Model for Group III

From Figure 31, we can see that there is a positive linear relationship between time and the two factors, age and duration. Although monotonic, it may well be linear. We also have quadric behavior between cigarettes per day and duration. Using this result, we can get a maximum model for Group 3 given by,

\[
Y_{g3} = \beta_0 + \beta_1 x_1^2 + \beta_2 x_1 + \beta_3 x_2^2 + \beta_4 x_2 + \beta_5 x_3 + \beta_6 x_3^2 + \sum_{i=7}^{n} \text{int}(x_1^2, x_1, x_2^2, x_2, x_3^2, x_3) + \epsilon
\]

(5.7)

Using backward selection, we have the final model in 5.8
\[ Y_{g3} = 12585.97 - 1041.6x_1 + 8300.3x_2 - 166.42x_3 - 50.3x_1^2 - 91.94x_2^2 - \\
598.07x_1 : x_2 + 11.66x_1 : x_3 - 124.55x_2 : x_3 + \\
2.41x_1^3 + 16.1x_2 : x_1^2 + 1.17x_3 : x_1^2 + 5.81x_1 : x_2^2 - \\
0.26x_3^3 + 1.88x_3 : x_2^2 - 0.14x_1^2 : x_2^2 + 8.34x_1 : x_2 : x_3 \] (5.8)

With multiple regression comes an overall test of significance, and a multiple $R^2$ which is actually the value of $r^2$ for the measured survival time vs. the predicted survival time. Our multiple R-squared is 0.9967, and adjusted R-squared is 0.9943 for our final Group 3 model which is Caucasian, former smoker, and female.

The following Table 17 shows us part of the residual validation.

Table 17: Residual Validation : Group III (Mean residual : -0.158 and standard error : 1.06).

<table>
<thead>
<tr>
<th>Actual Survival Time</th>
<th>Estimated Survival Time</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 809</td>
<td>786.26</td>
<td>-0.77</td>
</tr>
<tr>
<td>2 860</td>
<td>902.68</td>
<td>0.93</td>
</tr>
<tr>
<td>3 778</td>
<td>945.04</td>
<td>0.47</td>
</tr>
<tr>
<td>4 923</td>
<td>934.63</td>
<td>-1.08</td>
</tr>
<tr>
<td>5 879</td>
<td>781.50</td>
<td>-2.16</td>
</tr>
<tr>
<td>6 658</td>
<td>924.38</td>
<td>-0.72</td>
</tr>
<tr>
<td>7 859</td>
<td>807.94</td>
<td>1.04</td>
</tr>
<tr>
<td>8 782</td>
<td>841.23</td>
<td>0.55</td>
</tr>
<tr>
<td>9 868</td>
<td>905.67</td>
<td>-0.61</td>
</tr>
<tr>
<td>10 817</td>
<td>937.43</td>
<td>0.77</td>
</tr>
</tbody>
</table>
5.5.4 Multiple Regression Model for Group IV

Figure 32.: Matrix of Scatter Plots for Groups4.

From Figure 32, we can see that there is a positive linear relationship between time and two factors which are age and duration. Although monotonic, it well be linear. We also have quadric behavior between cigarettes per day and duration. Using this result, we can get a maximum model for Group4 given by,

$$Y_{g4} = \beta_0 + \beta_1 x_1^2 + \beta_2 x_1 + \beta_3 x_2^2 + \beta_4 x_2 + \beta_5 x_3 + \beta_6 x_3^2 + \sum_{i=7}^{n} int(x_1^2, x_1, x_2^2, x_2, x_3^2, x_3) + \epsilon$$

(5.9)

Using backward selection, we have the final model in 5.10
\[
Y_{s4}^* = 14.31 + 5.01x_1 + 8300.3x_2 + 14.1x_3 + 0.06x_1 : x_2 - \\
0.01x_1^3 + 0.02x_2 : x_2^2
\] (5.10)

With multiple regression comes an overall test of significance, and a multiple \( R^2 \) which is actually the value of \( r^2 \) for the measured survival time vs. the predicted survival time. Our multiple R-squared is 0.917, and adjusted R-squared is 0.888 for our final Group 4 model which is Caucasian, former smoker, and male.

The following Table 18 shows us part of the residual validation.

<table>
<thead>
<tr>
<th>Actual Survival Time</th>
<th>Estimated Survival Time</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  925</td>
<td>925.47</td>
<td>0.47</td>
</tr>
<tr>
<td>2  622</td>
<td>622.90</td>
<td>0.90</td>
</tr>
<tr>
<td>3  921</td>
<td>920.61</td>
<td>-0.39</td>
</tr>
<tr>
<td>4  880</td>
<td>878.06</td>
<td>-1.94</td>
</tr>
<tr>
<td>5  861</td>
<td>860.36</td>
<td>-0.64</td>
</tr>
<tr>
<td>6  902</td>
<td>902.32</td>
<td>0.32</td>
</tr>
<tr>
<td>7  862</td>
<td>861.26</td>
<td>-0.74</td>
</tr>
<tr>
<td>8  960</td>
<td>960.39</td>
<td>0.39</td>
</tr>
<tr>
<td>9  568</td>
<td>566.00</td>
<td>-2.00</td>
</tr>
<tr>
<td>10  783</td>
<td>782.14</td>
<td>-0.86</td>
</tr>
</tbody>
</table>

All four models provided high R-squares, ranging from 88% to 99%, which suggests that these models provide good predictions for survival times of lung cancer patients.
5.6 Fitting the Best Possible Probability Distribution

We proceeded to find the best possible probability distribution that characterizes the probabilistic behavior of the survival times for each group. Using three goodness of fit test methods, the Anderson-Darling Test(Anderson, Darling, 1951), Chi-Square Test(1954), and Kolmogorov Smirnov Test(Kolmogorov, Smirnov, 1971), we ranked all possible probability distributions using the results from the goodness of fit tests.

5.6.1 Goodness of Fit Test for Group All

On the Anderson goodness of fit test, the ranks of best fit from first to third rank is a Johnson SB probability distribution with four parameters, Generalized Gamma probability distribution with four parameters, and Beta probability distribution with four parameters. On the Chi-Squared Test, the ranks are Johnson SB probability distribution with four parameters, Power Function probability distribution, and Kumaraswamy probability distribution with four parameters. On the Komogorov Smirnov Test, the ranks are Generalized Gamma probability distribution with three parameters, Dagum probability distribution, and Wakeby probability distribution.

A summary of the rankings for the goodness-of-fit tests for Group I is shown in Table 19 below.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Anderson Darling Statistics</th>
<th>Rank</th>
<th>Distribution</th>
<th>Chi-Squared Statistics</th>
<th>Rank</th>
<th>Distribution</th>
<th>Kolmogorov Smirnov Statistics</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Johnson SB</td>
<td>5.1648</td>
<td>1</td>
<td>Johnson SB</td>
<td>75.286</td>
<td>1</td>
<td>Wakeby</td>
<td>0.00607</td>
<td>1</td>
</tr>
<tr>
<td>Gen.Gamma(4p)</td>
<td>6.7065</td>
<td>2</td>
<td>Beta</td>
<td>202.79</td>
<td>2</td>
<td>Johnson SB</td>
<td>0.00734</td>
<td>2</td>
</tr>
<tr>
<td>Beta</td>
<td>16.081</td>
<td>3</td>
<td>Kumaraswamy</td>
<td>162.32</td>
<td>3</td>
<td>G.. Gamma(4p)</td>
<td>0.00916</td>
<td>3</td>
</tr>
</tbody>
</table>

Evaluating the result of the goodness-of-fit tests is the four parameter Johnson SB probability distribution with MLE given by \((\gamma = -1.7143, \delta = 1.3379, \lambda = 668.27, \xi = 322.74)\).
\[ f(x) = \frac{\delta}{\lambda \sqrt{2\pi z(1-z)}} \exp\left(-\frac{1}{2} \left(\gamma + \delta \ln \left(\frac{z}{1-z}\right)\right)^2\right) \]  

(5.11)

where \( z = \frac{x-\xi}{\lambda} \), \( \gamma \) and \( \delta \) are shape parameters and \( \lambda \) is scale parameter \( \xi \) is location parameter, and the domain is \( \xi \leq x \leq \xi + \lambda \). The estimated form of the subject PDF is given by,

\[ f(t) = \frac{1.3379}{686.27 \sqrt{2\pi z(1-z)}} \exp\left(-\frac{1}{2} \left(-1.7143 + 1.3379 \ln \left(\frac{Z}{1-Z}\right)\right)^2\right), \quad Z \equiv \frac{t-322.74}{1.743} \]  

(5.12)

Thus, the parametric survival function for Group All is given by,

\[ S(t) = 1 - F(t) = 1 - \Phi \left(-1.7143 + 1.3379 \ln \left(\frac{Z}{1-Z}\right)\right), \quad Z \equiv \frac{t-322.74}{1.743} \]  

(5.13)

The graph of the survival function \( S(t) \) for the group all is given below, in Figure 33.

Figure 33.: Compare Survival Function with Parametric Method and K-M in Group All.
5.6.2 Goodness of Fit Tests for Group I

Based on the Anderson Darling goodness-of-fit test, the best is the Generalized extreme value probability distribution; the second best is the Log-Pearson probability distribution with three parameters; the third best is the Beta probability distribution with four parameters. Using the Chi-Squared Test, the first is the Generalized extreme value probability distribution; the second is the Log-Pearson probability distribution with three parameters; the third best is the Johnson SB probability distribution with four parameters. From the Komogorov Smirnov Test, the best is the Wakeby probability distribution; the second best is the Generalized Extreme value probability distribution; the third probability distribution is the Beta probability distribution.

Summary of the rankings of goodness-of-fit tests for Group I is given by Table 20 below,

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Anderson Darling Statistics</th>
<th>Rank</th>
<th>Chi-Squared Distribution Statistics</th>
<th>Rank</th>
<th>Kolmogorov Smirnov Distribution Statistics</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen. Ext. Value</td>
<td>0.0782</td>
<td>1</td>
<td>Gen. Ext. value</td>
<td>71.821</td>
<td>1</td>
<td>Wakeby</td>
</tr>
<tr>
<td>Log-Pearson 3</td>
<td>0.024</td>
<td>2</td>
<td>Log-Pearson 3</td>
<td>110.95</td>
<td>2</td>
<td>Gen. Ext. Value</td>
</tr>
<tr>
<td>Beta</td>
<td>0.021</td>
<td>3</td>
<td>Johnson SB</td>
<td>175.35</td>
<td>3</td>
<td>Beta</td>
</tr>
</tbody>
</table>

Evaluating the result of the goodness-of-fit tests is the Generalized extreme value probability distribution with MLE given by ($\hat{k} = -0.62429, \hat{\sigma} = 96.18$, and $\hat{\mu} = 827.85$). The analytical form of the subject PDF is given by,

$$f(t) = \begin{cases} \frac{1}{\sigma} \exp\left(-\left(1 + k z\right)^{-1/k}\right) \left(1 + k z\right)^{-1-1/k}, & k \neq 0 \\ \frac{1}{\sigma} \exp(-z - \exp(-z)), & k = 0, \end{cases}$$

(5.14)

where $z \equiv \frac{t-\mu}{\sigma}$, $k$ is continuous a shape parameter, $\sigma$ is a continuous scale parameters, and $\mu$ is a continuous location parameter. The estimated form of the subject PDF is given by,
\[
f_{1}(t) = \frac{1}{96.18} \exp\left(-\left(1 - 0.624 \frac{t - 827.85}{96.18}\right)^{1.603}\right) \left(1 - 0.624 \frac{t - 827.85}{96.18}\right)^{0.603},
\] (5.15)

Thus, the parametric survival function for Group I is given by,

\[
\hat{S}_{1}(t) = 1 - F_{2}(t) = 1 - \exp\left(-\left(1 - 0.624 \frac{t - 827.85}{96.18}\right)^{1.603}\right)
\] (5.16)

The graph of the survival function \( S_{1}(t) \) for Group I is given below, by Figure 34.

Figure 34.: Compare Survival Function with Parametric Method and K-M in Group I.

There is a significant difference after 600 months of survival time between overall survival curve and Group I.

5.6.3 Goodness of Fit Test for Group II

Based on the Anderson Darling goodness-of-fit test, the best is the Generalized extreme value probability distribution; the second best is the JohnsonSB probability distribution.
with four parameters; the third best is the Kumaraswamy probability distribution. Using the Chi-Squared Test, the first is the Generalized extreme value probability distribution; the second is the Johnson SB probability distribution with four parameters; the third best is Log-Pearson probability distribution with three parameters. From the Komogorov Smirnov Test, the best is the Wakeby probability distribution; the second best is the Pert probability distribution; the third best probability distribution is the Beta probability distribution.

Summary of rankings of goodness-of-fit test for Group II is given by Table 21, below,

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Anderson Darling Statistics</th>
<th>Rank</th>
<th>Chi-Squared Statistics</th>
<th>Rank</th>
<th>Kolmogorov Smirnov Statistics</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen. Ext. Value</td>
<td>0.0782</td>
<td>1</td>
<td>58.1</td>
<td>1</td>
<td>0.012</td>
<td>1</td>
</tr>
<tr>
<td>Johnson SB</td>
<td>0.027</td>
<td>2</td>
<td>106.13</td>
<td>2</td>
<td>0.018</td>
<td>2</td>
</tr>
<tr>
<td>Kumaraswamy</td>
<td>0.027</td>
<td>3</td>
<td>129.74</td>
<td>3</td>
<td>0.02</td>
<td>3</td>
</tr>
</tbody>
</table>

Evaluating the result of the goodness-of-fit tests is the Generalized extreme value probability distribution with MLE given by \(\hat{k} = -0.609, \hat{\sigma} = 98.448, \) and \(\hat{\mu} = 820.39\). The analytical form of the subject PDF is given by,

\[
f(t) = \begin{cases} 
\frac{1}{\sigma} \exp\left(-\left(1 + k z^{1/k}\right)\right)\left(1 + k z\right)^{-1 - 1/k}, & k \neq 0 \\
\frac{1}{\sigma} \exp(-z - \exp(-z)), & k = 0, 
\end{cases} \tag{5.17}
\]

where \(z \equiv \frac{t-\mu}{\sigma}\), \(k\) is continuous a shape parameter, \(\sigma\) is a continuous scale parameters, and \(\mu\) is a continuous location parameter. The estimated form of the subject PDF is given by,

\[
f_2(t) = \frac{1}{98.448} \exp\left(-\left(1 - 0.606 \frac{t-820.39}{98.448}\right)\right)\left(1 - 0.606 \frac{t-820.39}{98.448}\right)^{0.642}, \tag{5.18}
\]

Thus, the parametric survival function for Group II is given by,
\[ S_2(t) = 1 - F_2(t) = 1 - \exp \left( -1 - 0.606 \frac{t - 820.39}{98.448} \right)^{1.642} \] (5.19)

The graph of the survival function \( S_2(t) \) for Group II is given below, by Figure 35.

![Graph of survival function](image)

**Figure 35.:** Compare Survival Function with Parametric Method and K-M in Group II.

There is a significant difference in 600-920 months of survival time between overall survival curve and Group II.

### 5.6.4 Goodness of Fit Test for Group III

Based on the Anderson Darling goodness-of-fit test, the best is the Pert probability distribution; the second best is the Generalized extreme value probability distribution; the third best is the JohnsonSB probability distribution with four parameters. Using the Chi-Squared Test, the first is the Pert probability distribution; the second is the Beta probability distribution with four parameters; the third best is the Generalized extreme value probability distribution. From the Komogorov Smirnov Test, the best is the Wakeby probability distribution.
distribution; the second best is the Pert probability distribution; the third best probability
distribution is the Beta probability distribution.

Summary of rankings of goodness-of-fit test for Group III is given by Table 22, below,

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Anderson Darling Statistics</th>
<th>Rank</th>
<th>Chi-Squared Distribution Statistics</th>
<th>Rank</th>
<th>Kolmogorov Smirnov Distribution Statistics</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pert</td>
<td>38.846</td>
<td>1</td>
<td>Pert</td>
<td>160.2</td>
<td>Wakeby</td>
<td>0.016</td>
</tr>
<tr>
<td>Gen. Ext. Value</td>
<td>38.214</td>
<td>2</td>
<td>Beta</td>
<td>237.51</td>
<td>Pert</td>
<td>0.016</td>
</tr>
<tr>
<td>Johnson SB</td>
<td>0.0307</td>
<td>3</td>
<td>Gen. Ext. Value</td>
<td>247.52</td>
<td>Beta</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Evaluating the result of the goodness-of-fit tests is the pert probability distribution with
MLE given by \( \hat{m} = 924.66, \hat{a} = 417, \) and \( \hat{b} = 966 \). The analytical form of the subject PDF
is given by,

\[
f(t) = \frac{1}{B(\alpha_1, \alpha_2)} \frac{(t-a)^{\alpha_1-1}(b-t)^{\alpha_2-1}}{(b-a)^{\alpha_1+\alpha_2-1}},
\]

(5.20)

where \( B \) is the Beta Function which is

\[
B(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1-1}(1-t)^{\alpha_2-1}dt \quad (\alpha_1, \alpha_2 > 0),
\]

(5.21)

\( m \) (most likely value) - cuntinuous mode parameter \( (a \leq m \leq b) \), \( a \) and \( b \) are contunuous
boundary parameters \( (a < b) \), \( \alpha_1 = \frac{4m+b-5a}{b-a} = 4.7 \), and \( \alpha_2 = \frac{5b-a-4m}{b-a} = -5.13 \). The
estimated form of the subject PDF is given by,

\[
f_3^{\hat{}}(t) = \frac{1}{B(4.7,-5.13)} \frac{(t+417)^{3.7}(966-x)^{-6.13}}{(1383)^{-1.43}},
\]

(5.22)

Thus, the parametric survival function for Group III is given by,

\[
S_3^{\hat{}}(t) = 1 - F_3(t) = 1 - I_c(4.7,-5.13).
\]

(5.23)

70
where \( Z \equiv \frac{t-a}{b-a} \), \( B \) is the Beta Function, and \( I_z \) is the Regularized Incomplete Beta Function.

The graph of the survival function \( S_3(t) \) for Group III is given below, by Figure 36.

Figure 36.: Compare Survival Function with Parametric Method and K-M in Group III.

There is a significant difference after 760 months of survival time between the overall survival curve and Group III.

5.6.5 Goodness of Fit Test for Group IV

Based on the Anderson Darling goodness-of-fit test, the best is the Beta probability distribution with four parameters; the second best is the Generalized extreme value probability distribution; the third best is the Pert probability distribution. Using the Chi-Squared Test, the first is the Log-Person probability distribution with three parameters; the second is the Generalized extreme value probability distribution; the third best is the Beta probability distribution with four parameters probability distribution. From the Kolmogorov Smirnov Test, the best is the Beta probability distribution with four parameters; the second best is
the Wakeby probability distribution; the third best probability distribution is the Pert probability distribution.

Summary of rankings of goodness-of-fit test for Group IV is given by Table 23, below.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Anderson Darling Statistics</th>
<th>Rank</th>
<th>Distribution</th>
<th>Chi-Squared Statistics</th>
<th>Rank</th>
<th>Distribution</th>
<th>Kolmogorov Smirnov Statistics</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta</td>
<td>12.049</td>
<td>1</td>
<td>Log-Person 3</td>
<td>40.175</td>
<td>1</td>
<td>Beta</td>
<td>0.014</td>
<td>1</td>
</tr>
<tr>
<td>Pert</td>
<td>0.0169</td>
<td>3</td>
<td>Beta</td>
<td>44.966</td>
<td>3</td>
<td>Pert</td>
<td>0.017</td>
<td>3</td>
</tr>
</tbody>
</table>

Evaluating the result of the goodness-of-fit tests is the pert probability distribution with MLE given by \((\alpha_1 = 4.6829, \quad \alpha_2 = 1.1474, \quad a = 382, \quad \text{and} \quad b = 966)\),

\[
 f(t) = \frac{1}{B(\alpha_1, \alpha_2)} \frac{(t-a)^{\alpha_1-1}(b-t)^{\alpha_2-1}}{(b-a)^{\alpha_1+\alpha_2-1}},
\]

where \(B\) is the Beta Function which is

\[
 B(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1-1}(1-t)^{\alpha_2-1} dt \quad (\alpha_1, \alpha_2 > 0),
\]

\(\alpha_1\) and \(\alpha_2\) are shape parameters and \(a\) and \(b\) are boundary parameters \((a < b)\). The estimated form of the subject PDF is given by,

\[
 f_4(t) = \frac{1}{B(4.68, 1.15)} \frac{(t-382)^{3.68}(966-x)^{0.15}}{(584)^{4.83}},
\]

Thus, the parametric survival function for Group IV is given by,

\[
 S_4(t) = 1 - F_4(t) = 1 - I_z(4.68, 1.15).
\]

where \(Z \equiv \frac{t-a}{b-a}, \quad B\) is the Beta Function, and \(I_z\) is the Regularized Incomplete Beta Function.
The graph of the survival function $S_4(t)$ for Group IV is given below, by Figure 37.

![Survival Function Graph](image)

**Figure 37.** Compare Survival Function with Parametric Method and K-M in Group IV.

There is a significant difference after 720 months of survival time between the overall survival curve and Group IV.

The graph of the survival function for the groups and overall survival curve is given below, by Figure 38.

![Survival Curve Graph](image)

**Figure 38.** Compare Survival Curve Between Groups And Overall Survival Time.
The green curve shows the overall survival curve. The red and yellow curves are current smoking groups, and the purple and orange curves are former smoking groups. From the survival curve, there is a significant difference between current smoking groups and former smoking groups. Furthermore, the survival rate in current smoking groups is lower than the overall survival curve, and the survival rate of former smoking groups are higher than overall survival curve with current smoking groups.

5.7 Conclusion and Contributions

We proceeded to investigate the relationship between survival time and other attributable variables, such as number of cigarettes per day, duration, and age at diagnosis where a multiple regression model is the most commonly used tool. We applied several transformation methods and included polynomial terms; we obtained our final four models which provide a good R-square measure, which will be useful to predict survival times in lung cancer patients. From the results of the Goodness of fit tests, we ranked the 65 distributions using the Kolmogorov Smirnov, Anderson Darling, and Chi-Squared Test. We found the best possible probability distribution for each of the groups among the 65 distributions. There is a significant difference between groups and the overall survival curves.

Using our results, we could calculate a lung cancer patient’s survival time and probability information using number of cigarettes per day, duration, and age at diagnosis, by using our final models and probability density distributions.
Table 24: Summary of Mathematical Models, Probability Density and Survival Function.

<table>
<thead>
<tr>
<th>Group</th>
<th>Mathematical Functions</th>
</tr>
</thead>
</table>
| Group I     | $\hat{Y}_1 = -4202.01 + 527.82 \times x_1 - 159.37 \times x_2 + 84.43 \times x_3 - 16.92 \times x_1^2 + 3.82 \times x_2^2$
|             | $\begin{align*}
+ & 7.13 \times x_1 : x_2 - 8.15 \times x_1 : x_3 + 1.96 \times x_2 : x_3 \\
+ & 0.11 \times x_1^3 - 0.03 \times x_2 : x_1^2 + 0.21 \times x_3 : x_1^2 - 0.42 \times x_1 : x_2^2 \\
+ & 0.01 \times x_2^3 - 0.06 \times x_3 : x_2^2 + 0.01 \times x_1^2 : x_2^2 + 0.0018 \times x_1^3 : x_3
\end{align*}$
| $f_1(t) = \frac{1}{96.18} \exp \left(-\frac{t-827.85}{96.18}\right)^{1.603} (1 - \frac{0.624(t-827.85)}{96.18})^{0.603}$
| $S_1(t) = 1 - F_1(t) = 1 - \exp \left(-\frac{t-827.85}{96.18}\right)^{1.603}$ |
| Group II    | $\hat{Y}_2 = -9779.15 + 1898.28 \times x_1 + 772.32 \times x_2 + 187.12 \times x_3 - 114.36 \times x_1^2 - 6.06 \times x_2^2$
|             | $\begin{align*}
+ & 142.67 \times x_1 : x_2 - 33.16 \times x_1 : x_3 - 15.68 \times x_2 : x_3 + \\
+ & 2.2 \times x_1^3 + 8.42 \times x_2 : x_1^2 + 1.98 \times x_3 : x_1^2 + 1.1 \times x_1 : x_2^2
\end{align*}$
| $f_2(t) = \frac{1}{98.448} \exp \left(-\frac{t-820.39}{98.448}\right)^{1.642} (1 - \frac{0.606(t-820.39)}{98.448})^{0.642}$
| $S_2(t) = 1 - F_2(t) = 1 - \exp \left(-\frac{t-820.39}{98.448}\right)^{1.642}$ |
| Group III   | $\hat{Y}_3 = 12585.97 - 1041.6 \times x_1 + 8300.3 \times x_2 - 166.42 \times x_3 - 50.3 \times x_1^2 - 91.94 \times x_2^2$
|             | $\begin{align*}
+ & 598.07 \times x_1 : x_2 + 11.66 \times x_1 : x_3 - 124.55 \times x_2 : x_3 + \\
+ & 2.41 \times x_1^3 + 16.1 \times x_2 : x_1^2 + 1.17 \times x_3 : x_1^2 + 5.81 \times x_1 : x_2^2 \\
+ & 0.26 \times x_2^3 + 1.88 \times x_3 : x_2^2 - 0.14 \times x_1^2 : x_2^2 + 8.34 \times x_1 : x_2 : x_3
\end{align*}$
| $f_3(t) = \frac{1}{1583.13} \left(\frac{(t+417.57)(966-x)}{1383.13}\right)^{-1.13}$
| $S_3(t) = 1 - F_3(t) = 1 - L_2(4.7, -5.13)$ |
| Group IV    | $\hat{Y}_4 = 14.31 + 5.01 \times x_1 + 8300.3 \times x_2 - 14.1 \times x_3 - 0.06 \times x_1 : x_2$
|             | $\begin{align*}
+ & -0.01 \times x_1^3 + 0.02 \times x_2 : x_2^2
\end{align*}$
| $f_4(t) = \frac{1}{584.48} \left(\frac{(t-382.15)(966-x)}{584.48}\right)^{0.15}$
| $S_4(t) = 1 - F_4(t) = 1 - L_2(4.68, 1.15)$ |
| Group All   | $f(t) = \frac{1.3379}{686.27\sqrt{2\pi} \sigma(1-z)} \exp \left(-\frac{1}{2} \left(-1.7143 + 1.3379 \ln \left(\frac{z}{1-z}\right)\right)^2\right), \ Z \equiv t^{\frac{1}{1.7143}}$
| $\hat{S}(t) = 1 - F(t) = 1 - \Phi \left(-1.7143 + 1.3379 \left(\frac{Z}{1-Z}\right)\right), \ Z \equiv t^{\frac{1}{1.7143}}$ |
Figure 39.: Estimated Survival Time and Probability with Three factors.
Chapter 6
Future Research Projects

We introduced the four parameters logistic model in equation 2.1 and the five parameters logistic model in equation 2.3. We showed the graphical parameters and their effects on the five and four parameters logistic models. The four parameters logistic model is a special case in a five parameters logistic model when $g$ is 1. We can, however, apply the four parameter logistic model for symmetrical data. This proves that there are many restrictions in the four parameters logistic model because most data simply are not symmetrical. On the other hand, the proposed five parameters logistic model can handle asymmetrical medical data. There are two types of resistance: the first one being the parameter $B$ reaches 0 in the middle of the dosage. In this case, we cannot estimate the IC50, but we can estimate the ED50. The second type of resistance occurs when the estimated $B$ parameter is negative. In this case, we proceed to estimate the parameters using the three parameters logistic model, which is a special case in the four parameters logistic model when $B$ is 0. This program will give medical scientists scatter data plots, a fitted regression curve, and a result file which includes estimated parameters, indications of outliers, and indication of a need to continue experimentation. Medical scientists can reduce time spent on obtaining significant values for IC50 or ED50 and can easily view and compare results with new and other drug efficiencies. We also provided the excel result file generated by our Drug Efficiency Estimator GUI program. We propose to keep this program up to date and input new statistical functions on this software for medical scientists and researchers.

We also proposed to introduce new methodology using a randomized decision tree by utilizing bootstrapping methods to partition tree-based models to fit the SEER Breast Can-
cancer Dataset. Our proposed method is to rank the risk factors and to find the best classified points and risk factors for each node. We also propose to study a new pruning rule by minimizing the tree’s overall mean squared error, and a new labeling rule based on identifying subgroups of specific survival models. Using Breast Cancer data, we will be able to validate our proposed method using simulations and a internal cross validation procedure. When another cancer data set is received, we will conduct an external cross validation to check the effectiveness of our developed method.

We developed an R package for a four parameters JohnsonSB probability distribution. The JohnsonUSF package includes generating random numbers, calculating probabilities, and inverse probabilities, and estimating the four parameters for the JohnsonSB probability distribution. We propose to develope a process to generate a sequence of random samples, with a package that utilizes a loop type of processor that can simultaneously generate random samples and estimates of categorical behavior of the probability distribution of each sample. Presently, there is no method to efficiently obtain information for each of the random samples. The proposed development of each package will allow scientists to:

- Apply MCMC to develop random number generators to compare previous studies.
- Apply the E-M algorithm to obtain the parameter estimations, including censoring information.
- Develop a Bayesian model and apply the E-M algorithm to obtain bayesian parameter estimatis.

We propose to investigate the relationship between survival time and other attributable variables, such as number of cigarettes per day, duration, and age at diagnosis where a multiple nonlinear regression models will generate the survival times predict to these such factors. With the development of such nonlinear statistical model, we should be able to rank the target contributable variable to the smallest significant risk factor. Once we have developed and evaluated the accuracy of the proposed model using R and R adjusted and
residual analysis, we will proceed to use surface response analysis to determinate the values of the risk factors that will maximize the survival time of a cancer patient.
References


About the Author

Bong-Jin Choi got his undergraduate degree at Kyungnam University in South Korea and came to the United States for his graduate education at the University of South Florida. He obtained his masters degree in Statistics, and continued pursuing his Ph.D. degree with Dr. Chris Tsokos. He had internships at H. Lee Moffitt Cancer Center and Research Institute and Center for Urban Transportation Research during his graduate studies which led to a permanent position at the University of North Carolina at Chapel Hill, Lineberger Comprehensive Cancer Center.