2011

Statistical Models for Environmental and Health Sciences

Yong Xu
University of South Florida, xuyong007@hotmail.com

Follow this and additional works at: http://scholarcommons.usf.edu/etd
Part of the American Studies Commons, Biostatistics Commons, and the Environmental Sciences Commons

Scholar Commons Citation

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 Models for
Environmental and Health Sciences

by

Yong Xu

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

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

Date of Approval:
April 14, 2011

Keywords: Statistical modeling, Global warming, carbon dioxide, Breast cancer,
Power law process, differential equation, Cox proportional hazard model,
Kaplan-Meier

Copyright © 2011, Yong Xu
Dedication

I would like to dedicate my dissertation to my parents: my father Chengqun Xu and my mother Shuyun Gong. Without their support I can never make my study wish to be realized. It is their encouragements and love that make me to follow my heart and achieve of what I dream of.

I would also like to dedicate my dissertation to my advisor: Dr. Chris P. Tsokos. Dr. Tsokos show the light of my life career and lead my way on the study of Statistics. He taught every thing important to me not only the broad knowledge of Mathematics and Statistics but also the way of life and the right attitude toward science and research. He is my life mentor and I will continuously learn from him.
Acknowledgements

I want to thank for all my family members and all my friends. It is their love that supports me for so many years and makes this life a wonderful miracle to me.

I want to thank all faculties in the Mathematics and Statistics department at University of South Florida. Dr. Chris P. Tsokos had provided me so much useful guidance and help on my research and support me throughout the entire study time at USF. Dr. Wonkuk Kim helped me on many researches and I had so many wonderful classes with him. Dr. Kandethody M. Ramachandran helped me build the foundation of statistics and through his hard training and kind help I can gradually become an independent researcher. Dr. Gangaram Ladde shared with me only the way to master the materials in statistics but also the critical thinking abilities throughout different areas. Dr. Marcus Mcwaters helped me on all kinds of issues of study life at USF. It is them and other professors make the study life at USF a pleasant and unique experience for me.

I want to thank all staffs in the Mathematics and Statistics department at USF: Frances Johnston, Beverly DeVine-Hoffmeyer, Sarina Maldonado, Denise Marks and Mary Ann Wengyn. It is them that make this department more like a family instead of a school. I will never forget their kindness and the smile on their faces.
# Table of Contents

List of Tables  
List of Figures  
Abstract  

**Chapter 1 Introduction**  
1.1 Global Warming  
1.2 Breast Cancer  
1.3 Survival Analysis  

**Chapter 2 Modeling Carbon Dioxide with a System of Differential Equations**  
2.1 Background and Data  
2.2 Literature Review  
2.3 Methodology  
2.4 Conclusions  

**Chapter 3 Attributable Variables with Interactions That Contribute To Carbon Dioxide In The Atmosphere**  
3.1 Background and Data  
3.2 Methodology  
3.3 Validation of the Proposed Model  
3.4 Conclusions  

**Chapter 4 Probabilistic Comparison of Survival Analysis Models using Simulation and Cancer Data**  
4.1 Background and Data  
4.2 Literature Review  
4.3 Methodology  
4.3.1 Survival analysis using simulation  
4.3.2 Statistical model validation for uncensored cancer data  
4.3.3 Statistical model validation for censored data  
4.4 Conclusions  

**Chapter 5 Identify Attributable Variables and Interactions in Breast Cancer**  
5.1 Background and Data  
5.2 Methodology  
5.3 Validation of the Proposed Model  
5.4 Conclusions  

**Chapter 6 Power Law Process for Evaluating Stage I & II Ductal Breast Cancer Treatment**  
6.1 Background and Data  
6.2 Methodology  
6.3 Conclusions
<table>
<thead>
<tr>
<th>Chapter 7 Statistical Modeling of Breast Cancer Using Differential Equations</th>
<th>102</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1 Background and Data</td>
<td>102</td>
</tr>
<tr>
<td>7.2 Methodology</td>
<td>104</td>
</tr>
<tr>
<td>7.3 Conclusions</td>
<td>116</td>
</tr>
<tr>
<td>Chapter 8 Conclusions and Future Research</td>
<td>117</td>
</tr>
<tr>
<td>8.1 Conclusions</td>
<td>117</td>
</tr>
<tr>
<td>8.2 Future Research</td>
<td>122</td>
</tr>
<tr>
<td>8.2.1 Global Warming</td>
<td>122</td>
</tr>
<tr>
<td>8.2.2 Cancer Research</td>
<td>122</td>
</tr>
<tr>
<td>References</td>
<td>123</td>
</tr>
<tr>
<td>About the Author</td>
<td>End Page</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1 Statistical Evaluation Criteria ............................................. 14
Table 2.2 Residual Analysis ................................................................. 15
Table 2.3 Future Estimation of IROC of CO2 due to Gas Fuels ............... 15
Table 2.4 Statistical Evaluation Criteria of CO2 due to Gas Fuels .......... 17
Table 2.5 Residual Analysis of CO2 due to Gas Fuels .......................... 17
Table 2.6 Future Estimation of IROC of CO2 due to Liquid Fuels .......... 18
Table 2.7 Statistical Evaluation Criteria of CO2 due to Liquid Fuels ...... 20
Table 2.8 Residual Analysis of CO2 due to Liquid Fuels ....................... 20
Table 2.9 Future Estimation of IROC of CO2 Emission due to Gas Flaring 20
Table 2.10 Statistical Evaluation Criteria of CO2 due to Cement ........... 22
Table 2.11 Residual Analysis of CO2 due to Cement ......................... 23
Table 2.12 Future Estimation of IROC of CO2 due to Cement ............... 23
Table 2.13 Statistical Evaluation Criteria of CO2 due to Bunker .......... 24
Table 2.14 Residual Analysis CO2 due to Bunker ............................... 25
Table 2.15 Future Estimation of IROC of CO2 due to Bunker ............... 26
Table 2.16 Statistical Evaluation Criteria of CO2 Emissions ................. 28
Table 2.17 Residual Analysis of CO2 Emissions ................................. 28
Table 2.18 Future Estimation of IROC of Cumulative CO2 Emissions .... 28
Table 2.19 Statistical Evaluation Criteria of CO2 in the Atmospheres ...... 30
Table 2.20 Residual Analysis of CO2 in the Atmospheres ................... 31
Table 2.21 Future Estimation of IROC of Cumulative CO2 in the Atmospheres 31
Table 3.1 Box-Cox Transformation for Normality ............................... 36
Table 3.2 PRESS Statistics for Best Three Models .............................. 
Table 3.3 Statistical Evaluation Criteria
Table 3.4 Rank of Variable According to Contributions
Table 3.5 Residual Analysis
Table 3.6 Residual Analysis for Transformed Data
Table 3.7 Residual Analysis for Cross Validation
Table 4.1 The Most Used Kernel Densities
Table 4.2 The Most Used Optimal Bandwidth
Table 4.3 Residual Analysis
Table 4.4 Residual Analysis Repeat 1000 times
Table 4.5 The data set variables
Table 4.6 Cox model summary
Table 4.7 Cox model’s significance
Table 4.8 Residual analysis for uncensored data
Table 4.9 Residual analysis for censored data
Table 5.1 List of attributable variables
Table 5.2 Box-Cox Transformation for Normality for the Original Transformed Data
Table 5.3 PRESS Statistics for Best Three Models
Table 5.4 List of Attributable Variables
Table 5.5 Rank of Variable According to Contributions
Table 5.6 Residual Analysis for Original data and Cross Validation
Table 6.1 Parameter estimation for PLP
Table 7.1 Residual Analysis of Breast Cancer Tumor Size in Group I
Table 7.2 Residual Analysis of ROC of Breast Cancer Tumor Size in Group I
Table 7.3 Residual Analysis of Breast Cancer Tumor Size in Group II
Table 7.4 Residual Analysis of ROC of Breast Cancer Tumor Size in Group II
Table 7.5 Residual Analysis of Breast Cancer Tumor Size in Group III
Table 7.6 Residual Analysis of ROC of Breast Cancer Tumor Size in Group III
List of Figures

Figure 1.1 The Schematic Diagram of CO2 Emissions 2
Figure 1.2 The Schematic Diagram of CO2 in the Atmosphere 3
Figure 2.1 The Locations of Sampling of CO2 in the Atmosphere 10
Figure 2.2 CO2 due to Gas Fuels from 1959 to 2004 12
Figure 2.3 Model of CO2 due to Gas Fuels 13
Figure 2.4 IROC of CO2 due to Gas Fuels 13
Figure 2.5 CO2 due to Liquid Fuels 16
Figure 2.6 Model of CO2 due to Liquid Fuels 16
Figure 2.7 IROC of CO2 due to Liquid Fuels 17
Figure 2.8 CO2 due to Gas Flaring 18
Figure 2.9 Model of CO2 due to Gas Flaring 19
Figure 2.10 IROC of CO2 due to Gas Flaring 19
Figure 2.11 CO2 due to Cement 21
Figure 2.12 Model of CO2 due to Cement 21
Figure 2.13 IROC of CO2 due to Cement 22
Figure 2.14 CO2 due to Bunker 23
Figure 2.15 Model of CO2 due to Bunker 24
Figure 2.16 IROC of CO2 due to Bunker 25
Figure 2.17 Cumulative CO2 Emissions from 1950 to 2005 26
Figure 2.18 Model of Cumulative CO2 Emissions 27
Figure 2.19 IROC of Cumulative CO2 Emissions 27
Figure 2.20 Cumulative CO2 in the Atmospheres from 1950 to 2005 29
Figure 2.21 Model of Cumulative CO2 in the Atmospheres
Figure 2.22 IROC of Cumulative CO2 in the Atmospheres
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>QQ Plot for Testing Normality for Original Data</td>
<td>77</td>
</tr>
<tr>
<td>5.3</td>
<td>QQ Plot for Testing Normality for Transformed Data</td>
<td>78</td>
</tr>
<tr>
<td>5.4</td>
<td>Breast Cancer Attributable Variable Diagram</td>
<td>82</td>
</tr>
<tr>
<td>6.1</td>
<td>Breast Cancer Data Tree Diagram</td>
<td>89</td>
</tr>
<tr>
<td>6.2</td>
<td>Breast Cancer Data Tree Diagram white Ductal Stage I patients</td>
<td>90</td>
</tr>
<tr>
<td>6.3</td>
<td>Breast Cancer Data Tree Diagram white Ductal Stage II patients</td>
<td>91</td>
</tr>
<tr>
<td>6.4</td>
<td>Evaluation Chain for PLP</td>
<td>95</td>
</tr>
<tr>
<td>6.5</td>
<td>Stage I Breast Cancer Intensity Function without Treatment</td>
<td>96</td>
</tr>
<tr>
<td>6.6</td>
<td>Stage I Breast Cancer Intensity Function with Treatment 1</td>
<td>97</td>
</tr>
<tr>
<td>6.7</td>
<td>Stage I Breast Cancer Intensity Function with Treatment 2</td>
<td>97</td>
</tr>
<tr>
<td>6.8</td>
<td>Stage I Breast Cancer Intensity Function with Treatment 3</td>
<td>98</td>
</tr>
<tr>
<td>6.9</td>
<td>Stage II Breast Cancer Intensity Function without Treatment</td>
<td>99</td>
</tr>
<tr>
<td>6.10</td>
<td>Stage II Breast Cancer Intensity Function with Treatment 1</td>
<td>99</td>
</tr>
<tr>
<td>6.11</td>
<td>Stage II Breast Cancer Intensity Function with Treatment 2</td>
<td>100</td>
</tr>
<tr>
<td>6.12</td>
<td>Stage II Breast Cancer Intensity Function with Treatment 3</td>
<td>100</td>
</tr>
<tr>
<td>7.1</td>
<td>Breast Cancer Patients’ Tumor Size from Age 41 to Age 85</td>
<td>105</td>
</tr>
<tr>
<td>7.2</td>
<td>Breast Cancer Data Tree Diagram</td>
<td>105</td>
</tr>
<tr>
<td>7.3</td>
<td>Breast Cancer Patients’ Tumor Size from Age 41 to Age 58</td>
<td>106</td>
</tr>
<tr>
<td>7.4</td>
<td>Breast Cancer Patients’ Tumor Size from Age 41 to Age 58 with Curve Fitting</td>
<td>107</td>
</tr>
<tr>
<td>7.5</td>
<td>Breast Cancer Patients’ Tumor Size IROC from Age 41 to Age 58</td>
<td>109</td>
</tr>
<tr>
<td>7.6</td>
<td>Breast Cancer Patients’ Tumor Size from Age 59 to Age 73</td>
<td>110</td>
</tr>
<tr>
<td>7.7</td>
<td>Breast Cancer Patients’ Tumor Size from Age 59 to Age 73 with Curve Fitting</td>
<td>111</td>
</tr>
<tr>
<td>7.8</td>
<td>Breast Cancer Patients’ Tumor Size IROC from Age 59 to Age 73</td>
<td>112</td>
</tr>
<tr>
<td>7.9</td>
<td>Breast Cancer Patients’ Tumor Size from Age 73 to Age 85</td>
<td>113</td>
</tr>
<tr>
<td>7.10</td>
<td>Breast Cancer Patients’ Tumor Size from Age 74 to Age 85 with Curve Fitting</td>
<td>113</td>
</tr>
<tr>
<td>7.11</td>
<td>Breast Cancer Patients’ Tumor Size IROC from Age 73 to Age 85</td>
<td>113</td>
</tr>
</tbody>
</table>
Abstract

Statistical analysis and modeling are useful for understanding the behavior of different phenomena. In this study we will focus on two areas of applications: Global warming and cancer research. Global Warming is one of the major environmental challenge people face nowadays and cancer is one of the major health problem that people need to solve.

For Global Warming, we are interest to do research on two major contributable variables: Carbon dioxide (CO2) and atmosphere temperature. We will model carbon dioxide in the atmosphere data with a system of differential equations. We will develop a differential equation for each of six attributable variables that constitute CO2 in the atmosphere and a differential system of CO2 in the atmosphere. We are using real historical data on the subject phenomenon to develop the analytical form of the equations. We will evaluate the quality of the developed model by utilizing a retrofitting process. Having such an analytical system, we can obtain good estimates of the rate of change of CO2 in the atmosphere, individually and cumulatively as a function of time for near and far target times. Such information is quite useful in strategic planning of the subject matter. We will develop a statistical model taking into consideration all the attributable variables that have been identified and their corresponding response of the amount of CO2 in the atmosphere in the continental United States. The development of the statistical model that includes interactions and higher order entities, in addition to individual contributions to CO2 in the atmosphere, are included in the present study. The proposed model has been statistically evaluated and produces accurate predictions for a given set of the attributable variables. Furthermore, we rank the attributable variables with respect to their significant contribution to CO2 in the atmosphere.
For Cancer Research, the object of the study is to probabilistically evaluate commonly used methods to perform survival analysis of medical patients. Our study includes evaluation of parametric, semi-parametric and nonparametric analysis of probability survival models. We will evaluate the popular Kaplan-Meier (KM), the Cox Proportional Hazard (Cox PH), and Kernel density (KD) models using both Monte Carlo simulation and using actual breast cancer data. The first part of the evaluation will be based on how these methods measure up to parametric analysis and the second part using actual cancer data. As expected, the parametric survival analysis when applicable gives the best results followed by the not commonly used nonparametric Kernel density approach for both evaluations using simulation and actual cancer data. We will develop a statistical model for breast cancer tumor size prediction for United States patients based on real uncensored data. When we simulate breast cancer tumor size, most of time these tumor sizes are randomly generated. We want to construct a statistical model to generate these tumor sizes as close as possible to the real patients’ data given other related information. We accomplish the objective by developing a high quality statistical model that identifies the significant attributable variables and interactions. We rank these contributing entities according to their percentage contribution to breast cancer tumor growth. This proposed statistical model can also be used to conduct surface response analysis to identify the necessary restrictions on the significant attributable variables and their interactions to minimize the size of the breast tumor.

We will utilize the Power Law process, also known as Non-homogenous Poisson Process and Weibull Process to evaluate the effectiveness of a given treatment for Stage I & II Ductal breast cancer patients. We utilize the shape parameter of the intensity function to evaluate the behavior of a given treatment with respect to its effectiveness. We will develop a differential equation that will characterize the behavior of the tumor as a function of time. Having such a
differential equation, the solution of which once plotted will identify the rate of change of tumor size as a function of age. The structure of the differential equation consists of the significant attributable variables and their interactions to the growth of breast cancer tumor. Once we have developed the differential equations and its solution, we proceed to validate the quality of the proposed differential equations and its usefulness.
Chapter 1

Introduction

1.1 Global Warming

According to Wikipedia, Global warming is the increase in the average temperature of Earth’s near-surface air and oceans since the mid-20th century and its projected continuation. Nowadays, people have more and more awareness of global warming issue and want to contribute to handle this common challenge to all human kind on our planet. Therefore to understand Global Warming become a key issue in front of us.

Carbon dioxide emissions, CO2, along with atmospheric temperature are two of the key entities that contribute GLOBAL WARMING. In the present study we will be concerned with the six attributable variables that constitute CO2 emissions, namely, Gas fuels (G), Solid fuels (S), Liquid fuels (L), Gas Flares (F), Cement (C) and Bunker (B). A schematic diagram of CO2 emissions is given below, Figure 1.1.

CO2 in the atmosphere, along with atmospheric temperature, are two of the key entities that contribute to GLOBAL WARMING. In the present study we will be concerned with the all possible attributable variables that constitute CO2 in the atmosphere, namely, CO2 emission, Flux from Atmosphere to Oceans, Flux from Oceans to Atmosphere, Terrestrial Photosynthesis, Respiration, Burial of Organic Carbon & Limestone Carbon and Deforestation & Destruction. A schematic diagram of CO2 in the atmosphere is given below, Figure 1.2.

The eight attributable variables are namely, CO2 emission (E), deforestation and destruction of biomass and soil carbon (D), terrestrial plant respiration (R),
respiration from soils and decomposers (S), the flux from oceans to atmosphere (O), terrestrial photosynthesis (P), the flux from atmosphere to oceans (I), the burial of organic carbon and limestone carbon in sediments and soils (B).

![SCHEMATIC DIAGRAM FOR CO2 EMISSION (2005)](image)

**Figure 1.1 The Schematic Diagram of CO2 Emissions**

We need to mention here that some of the attributable variables are the function of several other variables within themselves. For example, CO2 emission, E, is a function of six attributable variables namely, Gas fuels (Ga), Solid fuels (So), Liquid fuels (Li), Gas Flares (Fl), Cement (Ce) and Bunker (Bu). Gas fuels include gas consisting primarily of methane. They include natural gas and other gases that can provide energy through combustion. Solid fuels refer to various types of solid material that are used as fuel to produce energy and provide heating, usually released through combustion. Solid fuels include wood, charcoal, coal and others. Liquid fuels are those combustible or energy-generating molecules that can be harnessed to create mechanical energy, such as the gasoline we normally use. Gas flares is the vertical stack on oil wells or natural gas well completion activities.
Cement refers to the Co2 generated through the production of cement. Bunker fuel is a type of crude oil also named heavy oil or furnace oil. It belongs to the heavy fractions or hard to distill fractions when crude oil is refined and often used for ships.

![Image](image.png)

**Figure 1.2 The Schematic Diagram of CO2 in the Atmosphere**

### 1.2 Breast Cancer

According to Wikipedia, Cancer (medical term: malignant neoplasm) is a class of diseases in which a cell, or a group of cells display uncontrolled growth (division beyond the normal limits), invasion (intrusion on and destruction of adjacent tissues), and sometimes metastasis (spread to other locations in the body via lymph or blood). These three malignant properties of cancers differentiate them from benign
tumors, which are self-limited, and do not invade or metastasize. Most cancers form a tumor but some, like leukemia, do not.

Scientists believe that cancer is related with people’s age, lifestyle, environment and genetics. In our study, we mainly focus on breast cancer. The definition of breast cancer is: breast cancer (malignant breast neoplasm) is cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk.

According to National Cancer Institute (NCI)' Surveillance Epidemiology and End Results (SEER) results, it is estimated that 207,090 women will be diagnosed with and 39840 women will die of breast cancer in 2010.

From 2003 to 2007, the median age at diagnosis for breast cancer was 61 years of age. The highest rate age group is 55 to 64 with 24.1% followed by 45 to 54 with 22.6% and approximately 0% for under age 20 as smallest rate group. The age-adjusted incidence rate was 122.9 per 100,000 women per year. All these rates are based on cases from 17 SEER 2003-2007. White female has the highest incidence rate with 126.5 per 100,000, followed by Black with 118.3 per 100,000, Asian 90 per 100,000, Hispanic 86 per 100,000 and American Indian 76.4 per 100,000.

According to SEER patients who died in 2003 to 2007, the median age at death for breast cancer was 68 years old. The highest age group is 75 to 84 with 22.6%, followed by 55 to 64 with 20.8% and almost 9% for under age 20 as the lowest. Black female have the highest death rate 32.4 per 100,000, followed by White 23.4 per 100,000, American Indian 17.6 per 100,000, Hispanic 15.3 per 100,000 and Asian 12.2 per 100,000. The age-adjusted death rate was 24.0 per 100,000 women per year.

The trend in SEER cancer incidence between 1975 and 2007 shows that the highest increase is from 1980 to 1987 with 3.9, followed by from 1995 to 1998 with 2.7 and the smallest increase is from 1998 to 2007 with negative 1.7.
From 17 SEER data, the overall 5-year relative survival was 89%. Five-year survival was: 90.2% for white women; 77.5% for black women. Localized stage has the highest 5-year relative survival with 98%, followed by regional stage with 83.6%, unstaged 57.9% and distant stage with 23.4%. As for lifetime risk, 12.15% of women born today will be diagnosed with breast cancer at some time during their lifetime based on rates from 2005 to 2007.

1.3 Survival Analysis

According to Wikipedia, Survival analysis is a branch of statistics which deals with death in biological organisms and failure in mechanical systems. This topic is called reliability theory or reliability analysis in engineering, and duration analysis or duration modeling in economics or sociology. More generally, survival analysis involves the modeling of time to event data; in this context, death or failure is considered an "event" in the survival analysis literature. Many concepts in Survival analysis have been explained by the Counting Process Theory, which has emerged more recently. The flexibility of a counting process is that it allows modeling multiple (or recurrent) events.

When we apply survival analysis for biological problem such as cancer problem, we will deal with missing information of the death of the patients due to many reasons such as patients transfer to other hospital, patients died because of other diseases or accidents, the research project meet the deadline, etc. Therefore, we need to consider censoring. Censoring simply means we only have part of the information. For example, if one patient is transfer to other hospital and decide to change the medicine after one year, we can only record as the patient is still alive up to one year after the medicine is applied. We do not know exactly when this patient will be dead but only have partial information that this patient live at least one year.
We have several types of censoring. Left censoring refer to the information or data point is below a certain value and we do not know how much it is. Right censoring means the information or data point is above a certain value and we do not know how much it is. The previous example will belong to the right censoring case since we only know it will be at least one year and do not know exactly how long. The interval censoring belongs to the situation that the information or data point is between two values and it looks like left and right censoring cases together. There are also three types of censorings. Type I censoring defined as if a trail has a set number of subjects and stops the trail at a predetermined time. Therefore any subjects remaining (still in survival) will be considered as right censored case and all others will be uncensored case since we do have full information about them. Type II censoring occurs if a trail has a set number of subjects and stops the trail when a predetermined number of subjects have failed. Like type I censoring, the remaining subject are all right censored. Random censoring, sometimes called non-informative censoring, occurs when each subject has its own censoring time that is independent of its failure time. Therefore, the time we observed is the minimum of the random censoring and the actual failure time. For those subjects whose failure time is longer than the censoring time are right censored.

Survival function also called survivorship function, normally denoted in Capital letter S, below.

\[ S(x) = P(X > x) = 1 \quad P(X \geq x) = 1 \quad F(x) \]  

(1.1)

where \( x \) is some time and will be a real number, \( X \) is a random variable stands for the time of dease. \( P \) denote for probability function and \( F \) is the cumulative probability function (CDF) also named as distribution function (DF) or lifetime distribution function. Since CDF stands for the probability for a cumulative time of dease less than or equal to some specified time, the survival function \( S \) will be the probability that the time of dease is longer than some specified time. In
another words, S stands for the chance for the patient to live more than the specific
time. Normally, we assume all patients are still live when we start our research.
Therefore, we will assume S(0)=1 means the probability of any one can live at the
beginning of the trail is one. If there are some immediate decease exits, S(0) can
also less than one. Since F(x) is non-decreasing function, S(x) must be non-
increasing. In another words, if a patient can live in later time, this same patient has
to be alive for now. S(x) normally will approach to zero as time increases to infinite if
we do not assume eternal life exist. The survival function is also called survivor
function in biological survival and reliability function in engineering survival.

F(x) is defined as the complement of the survival function as mentioned
above. The derivative of the F is denoted by f with the name probability density
function (pdf).

\[ f(x) = F'(x) = \frac{d}{dx} F(x) = \frac{d}{dx} (1 - S(x)) = S'(x) \tag{1.2} \]

f(x) stands for the rate of the decease and if f(x) is continuous, S(x) can be
written as

\[ S(x) = P(X > x) = \int_x^\infty f(t)dt \tag{1.3} \]

The hazard function \( h(x) \) is defined as the event at specified time conditional on
this event survival until the same specified time or later. The definition of \( h(x) \) is given
below,

\[ h(x) = \frac{f(x)}{S(x)} \]. \tag{1.4} \]

\[ h(x)dx = P(x < x + dx | X > x) = \frac{f(x)dx}{S(x)} \tag{1.5} \]

The hazard function has to be non negative since both f(x) and S(x) are non
negative. The cumulative hazard function, denoted by \( H(x) \) is defined below,
\( (x) = \int_0^x (x) dx = \int_0^x f(x) \frac{dx}{S(x)} = \int_0^x S'(x) \frac{dx}{S(x)} = \int_0^x \frac{S'(x)}{S(x)} dx = \ln(S(x)) \) (1.6)

And \( S(x) = \exp(\ln(x)) \). Since \( (x) = \ln(S(x)) \), we can see \( (x) \) increase without bound as \( x \) approaches to infinity which means \( (x) \) has to diverge.
Chapter 2

Modeling Carbon Dioxide with a System of Differential Equations

2.1 Background and Data

The object of the present study is to develop differential equations for each of the attributable variables of CO2 in the atmosphere, CO2 emission and the total CO2 in the atmosphere.

The data was obtained from Carbon Dioxide Information Analysis Center (CDIAC). The CDIAC is the primary climate-change data and information analysis center of the U.S. Department of Energy (DOE), located at Oak Ridge National Laboratory (ORNL) and includes the World Data Center for Atmospheric Trace Gases. All the individual emissions estimates are expressed in thousand metric tons of carbon (MT) and total CO2 in the atmosphere is in unit parts per million (PPM). There are ten different locations used to gather the samples of CO2 in the atmosphere. The ten locations are listed in Figure 2.1, below. ORNL recommends using Mauna Loa data because of the ocean current’s moving effect making the Mauna Loa data more representative. Therefore, we will use Mauna Loa’s CO2 in the atmosphere data. Carbon emissions are calculated by the fuels consumed times heat coefficient times the carbon coefficient times the combustion efficiency. The product of fuels consumed times heat coefficient is in the unit of trillion Btu. The carbon coefficients are given by the Environmental Protection Agency (EPA) reports (Blasing et. al. 2005). It is the amount of carbon that is emitted per unit of heat realized from combustion. Petroleum data come from the DOE reports, which are published in Monthly Energy Review (Blasing et. al. 2005, 2007) (Marland et. al 2007).
We need to mention here that some of the attributable variables are the function of several other variables within themselves. For example, CO2 emission, E, is a function of six attributable variables namely, Gas fuels (Ga), Solid fuels (So), Liquid fuels (Li), Gas Flares (Fl), Cement (Ce) and Bunker (Bu). Gas fuels include gas consisting primarily of methane. They include natural gas and other gases that can provide energy through combustion. Solid fuels refer to various types of solid material that are used as fuel to produce energy and provide heating, usually released through combustion. Solid fuels include wood, charcoal, coal and others. Liquid fuels are those combustible or energy-generating molecules that can be harnessed to create mechanical energy, such as the gasoline we normally use. Gas flares is the vertical stack on oil wells or natural gas well completion activities. Cement refers to the Co2 generated through the production of cement. Bunker fuel is a type of crude oil also named heavy oil or furnace oil. It belongs to the heavy fractions or hard to distill fractions when crude oil is refined and often used for ships.

![Map of CO2 Locations](image_url)

Figure 2.1 The Locations of Sampling of CO2 in the Atmosphere

**2.2 Literature Review**

Hansen (1984) discuss the climate processes and climate sensitivity. Ramanathan (1988) present the greenhouse theory of climate change. Lashof (198...
analyze the feedback processes that may influence future concentrations of atmospheric trace gases and climate change. Thomas J. Goreau (1990) stated the eight attributable variables for CO2 in the atmosphere. Retallack (2002) generally talk about the understanding climate change. Tsokos and Xu (2009) have proposed differential equations for individual attributable variables for CO2 emission and cumulatively. The parametric analysis for CO2 has been studied extensively by Wooten and Tsokos (2010). They have found that the CO2 data follows the three parameter Weibull probability distribution contrary to the fact that some scientists believed that CO2 in the atmosphere follows Gaussian probability distribution. We will use the most updated data to construct the individual differential equation systems for CO2 in the atmosphere, individually and cumulatively. In this study we will follow up to construct the differential equations for those significant attributable variables and cumulatively for CO2 in the atmosphere.

2.3 Methodology

Thomas J. Goreau (1990) briefly mentioned that the rate of change of CO2 in the atmosphere and also CO2 in the atmosphere should be studied using differential equations. However, to our knowledge no actual differential equations have been developed on the subject matter to study the rate of change of CO2. .

According to Chapter 2 result, we know only five of the six attributable variables do contribute to CO2 in the atmosphere therefore we will focus on only the five attributable variables namely, Gas fuels (Ga), Liquid fuels (Li), Gas Flares (Fl), Cement (Ce) and Bunker (Bu). Using A to represent CO2 in the atmosphere and the systematic representation of the five attributable variables, the functional form of the differential equations is given by equation 2.1 below.

\[
\frac{d(A)}{dt} = f\left(\frac{d(Ga)}{dt}, \frac{d(Li)}{dt}, \frac{d(Fl)}{dt}, \frac{d(Ce)}{dt}, \frac{d(Bu)}{dt}\right)
\]
Without loss of generality, we can express the rate of change of CO2 in the atmosphere as a linear function of all five attributable variables by

\[
\frac{d(A)}{dt} = C_1 \frac{d(Ga)}{dt} + C_2 \frac{d(Li)}{dt} + C_3 \frac{d(Fl)}{dt} + C_4 \frac{d(Ce)}{dt} + C_5 \frac{d(Bu)}{dt} + C_6 \quad (2.2)
\]

where C1 to C5 are the coefficients of each differential term and C6 is a constant. We shall begin to formulate the differential equations of each attributable variable.

CO2 in the atmosphere due to gas fuels. For CO2 in the atmosphere due to gas fuels, the data in the unit of metric tons is graphically shown by Figure 2.2, below.

The differential equation for gas fuels is given by

\[
G(x) + G'(x) = 9.551 \times 10^8 \times 10^{10} + 1.447 \times 10^8 x + 7.303 \times 10^4 x^2 + 1.229 \times 10^1 x^3 \quad (2.3)
\]

The x represents years in the above equation (2.3). The solution of 2.3 is given by

\[
G(x) = 9.566 \times 10^8 + 1.448 \times 10^8 x + 7.307 \times 10^4 x^2 + 1.229 \times 10^1 x^3 \quad (2.4)
\]

Figure 2.2 CO2 due to Gas Fuels from 1959 to 2004
A graphical display of the actual data and the solution of the differential equation for gas fuels is given by Figure 2.3.

![Figure 2.3 Model of CO2 due to Gas Fuels](image)

The instantaneous rate of change (IROC) of gas fuels as a function of time is given analytically by

\[
G'(x) = 1.448 \times 10^8 + 1.461 \times 10^5 x + 3.687 \times 10^1 x^2
\]

(2.5)

A graphical display of expression 2.5 is given by Figure 2.4.

![Figure 2.4 IROC of CO2 due to Gas Fuels](image)
Thus, one can utilize either equation 2.5 or the above graph to obtain the estimate of the rate of change for CO2 due to gas fuels for short and long terms of time. The question is how good are these estimates? The answer depends on the quality of the developed analytical models using the raw data. To test for the quality of the proposed analytical models, we use three statistical criteria, the $R^2$ ($R^2$ adjusted), the PRESS statistic and residual analysis.

The regression sum of squares (SSR), also called the explained sum of squares, is the variation that is explained by the regression model. The sum of squared errors (SSE), also called the residual sum of squares, is the variation that is left unexplained. The total sum of squares (SST) is proportional to the sample variance and equals the sum of SSR and SSE. The coefficient of determination $R^2$ is defined as the proportion of the total response variation that is explained by the model. It provides an overall measure of how well the model fits. $R^2$ adjusted will adjust for degree of freedom of the model and it works better when we have a lot of parameters. The prediction of residual error sum of squares (PRESS) statistics will evaluate how good the estimation will be if each time we remove one data (Allen 1971 and 1974).

The calculated values for the gas fuel model for $R^2$ ($R^2$ adjusted) and PRESS statistic are given by Table 2.1, below.

<table>
<thead>
<tr>
<th>R square</th>
<th>R square adjusted</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6676</td>
<td>0.6439</td>
<td>40597493062</td>
</tr>
</tbody>
</table>

The values of $R^2$ ($R^2$ adjusted) reflect the fact that we have identified a good model along with a PRESS statistic value that is the smallest of several models that we tested. Furthermore, the residual analysis we performed on the proposed differential equation of gas fuels is given in Table 2.2, below.
Table 2.2 Residual Analysis of CO2 due to Gas Fuels

<table>
<thead>
<tr>
<th>Year</th>
<th>Empirical ROC</th>
<th>DF IROC</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>0.029913</td>
<td>0.015397</td>
<td>0.014517</td>
</tr>
<tr>
<td>1996</td>
<td>0.000789</td>
<td>0.018589</td>
<td>-0.0178</td>
</tr>
<tr>
<td>1997</td>
<td>0.021466</td>
<td>0.021868</td>
<td>-0.0004</td>
</tr>
<tr>
<td>1998</td>
<td>-0.01184</td>
<td>0.025191</td>
<td>-0.03703</td>
</tr>
<tr>
<td>1999</td>
<td>0.02542</td>
<td>0.028514</td>
<td>-0.00309</td>
</tr>
<tr>
<td>2000</td>
<td>0.037914</td>
<td>0.031793</td>
<td>0.006121</td>
</tr>
<tr>
<td>2001</td>
<td>-0.07094</td>
<td>0.034987</td>
<td>-0.10593</td>
</tr>
<tr>
<td>2002</td>
<td>0.040557</td>
<td>0.03806</td>
<td>0.002497</td>
</tr>
<tr>
<td>2003</td>
<td>-0.03233</td>
<td>0.040977</td>
<td>-0.0733</td>
</tr>
<tr>
<td>2004</td>
<td>-0.01453</td>
<td>0.043713</td>
<td>-0.05824</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean of residual</th>
<th>Standard deviation of residual (SD)</th>
<th>Standard error of residual (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.02726648</td>
<td>0.04014411</td>
<td>0.01269468</td>
</tr>
</tbody>
</table>

Here, the empirical rate of change, Empirical ROC, is calculated using the actual data of gas fuels that we refer to as the true values and the instantaneous rate of change using the developed differential equation, DF IROC, with the residual being the difference of the two. As seen from the table, the residuals are extremely small and so is the standard error. These results attest to the good quality of the proposed model for gas fuels. Thus, in Table 2.3, below, we have calculated 10, 20 and 50 years ahead the instantaneous rate of change of gas fuel emissions.

Table 2.3 Future Estimation of IROC of CO2 due to Gas Fuels

<table>
<thead>
<tr>
<th>Years</th>
<th>10 years</th>
<th>20 years</th>
<th>50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IROC in future</td>
<td>36519.9</td>
<td>63922.39</td>
<td>190371.4</td>
</tr>
</tbody>
</table>

CO2 in the atmosphere due to liquid fuels. For liquid fuels the data in metric tons is graphically shown by Figure 2.5 below.

The differential equation for liquid fuels is given by

\[ L(x) + L'(x) = 1.624 \times 10^{11} + 2.455 \times 10^8 x + 1.237 \times 10^5 x^2 + 2.078 \times 10^1 x^3. \]  

(2.6)

The solution of (2.6) is given by

\[ L(x) = 1.626 \times 10^{11} + 2.458 \times 10^8 x + 1.238 \times 10^5 x^2 + 2.078 \times 10^1 x^3. \]  

(2.7)
A graphical display of the actual data and the solution of the differential equation for liquid fuels is given by Figure 2.6.

The IROC of liquid fuels as a function of time is given analytically by

\[ L'(x) = 2.456 \times 10^8 + 2.476 \times 10^5 x + 6.235 \times 10^1 x^2. \]  

(2.8)

Figure 2.5 CO2 due to Liquid Fuels

Figure 2.6 Model of CO2 due to Liquid Fuels

A graphical display of expression 2.8 is given by Figure 2.7.
Figure 2.7 IROC of CO2 due to Liquid Fuels

The calculated values for the liquid fuel model for $R^2$ ($R^2_{\text{adjusted}}$) and PRESS statistic are given by Table 2.4, below.

<table>
<thead>
<tr>
<th>R square</th>
<th>R square adjusted</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8898</td>
<td>0.7984</td>
<td>97414862717</td>
</tr>
</tbody>
</table>

The values of $R^2$ ($R^2_{\text{adjusted}}$) reflect that we have identified a good model along with a PRESS statistic value that is the smallest of several models that we tested. Furthermore, the residual analysis we performed on the proposed differential equation of gas fuels is given in Table 2.5, below.

Table 2.5 Residual Analysis of CO2 due to Liquid Fuels

<table>
<thead>
<tr>
<th>Year</th>
<th>Empirical ROC</th>
<th>DF ROC</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>-0.02197</td>
<td>0.005001</td>
<td>-0.02697</td>
</tr>
<tr>
<td>1996</td>
<td>0.008361</td>
<td>0.007056</td>
<td>0.001305</td>
</tr>
<tr>
<td>1997</td>
<td>0.010318</td>
<td>0.009285</td>
<td>0.001103</td>
</tr>
<tr>
<td>1998</td>
<td>0.021686</td>
<td>0.011669</td>
<td>0.010017</td>
</tr>
<tr>
<td>1999</td>
<td>0.01399</td>
<td>0.014184</td>
<td>-0.00019</td>
</tr>
<tr>
<td>2000</td>
<td>0.023859</td>
<td>0.016805</td>
<td>0.007054</td>
</tr>
<tr>
<td>2001</td>
<td>0.004259</td>
<td>0.019503</td>
<td>-0.01524</td>
</tr>
<tr>
<td>2002</td>
<td>-0.00426</td>
<td>0.022247</td>
<td>-0.02651</td>
</tr>
<tr>
<td>2003</td>
<td>0.030241</td>
<td>0.025007</td>
<td>0.005234</td>
</tr>
<tr>
<td>2004</td>
<td>0.021984</td>
<td>0.027752</td>
<td>-0.00577</td>
</tr>
</tbody>
</table>

Mean of residual: -0.00500436

Standard deviation of residual (SD): 0.01344009

Standard error of residual (SE): 0.004250129
Here, the empirical rate of change, Empirical ROC, is calculated using the actual data of liquid fuels that we refer to as the true values and the instantaneous rate of change using the developed differential equation, DF IROC, with the residual being the difference of the two. As seen from the table, the residuals are extremely small and so is the standard error. These results attest to the good quality of the proposed model for gas fuels. Thus, in Table 2.2.3, below, we have calculated 10, 20 and 50 years ahead the instantaneous rate of change of liquid fuel emissions.

<table>
<thead>
<tr>
<th>Years</th>
<th>10 years</th>
<th>20 years</th>
<th>50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IROC in future</td>
<td>48886.24</td>
<td>90873.68</td>
<td>291661.3</td>
</tr>
</tbody>
</table>

CO2 in the atmosphere due to gas flaring. For gas flaring the data in metric tons is graphically shown by Figure 2.8 below.

![Figure 2.8 CO2 due to Gas Flaring](image)

The differential equation for gas flaring is given by

\[ F(x) + F'(x) = 2.337 \times 10^7 + 2.347 \times 10^4 x + 5.895 x^2. \]  
(2.12)

The solution of 2.12 is given by

\[ F(x) = 2.339 \times 10^7 + 2.348 \times 10^4 x + 5.895 x^2. \]  
(2.)
A graphical display of the actual data and the solution of the differential equation for gas flaring is given by Figure 2.9.

![Figure 2.9 Model of CO2 due to Gas Flaring](image)

The IROC of gas fuel as a function of time is given analytically by

\[ F'(x) = 2.348 \times 10^4 + 11.791 \times x \]  

(2.14)

A graphical display of expression 2.14 is given by Figure 2.10.

![Figure 2.10 IROC of CO2 due to Gas Flaring](image)
The calculated values for the solid fuels model for $R^2$ ($R^2$ adjusted) and PRESS statistic are given by Table 2.7, below.

Table 2.7 Statistical Evaluation Criteria of CO2 due to Gas Flaring

<table>
<thead>
<tr>
<th>R square</th>
<th>R square adjusted</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7025</td>
<td>0.6886</td>
<td>75291841</td>
</tr>
</tbody>
</table>

The values of $R^2$ ($R^2$ adjusted) reflect the fact that we have identified a good model along with a PRESS statistic value that is the smallest of several models that we tested. Furthermore, the residual analysis we performed on the proposed differential equation of gas flaring is given in Table 2.8, below.

As seen from the table, the residuals are extremely small and so is the standard error. These results attest to the good quality of the proposed model for gas flaring. Thus, in Table 2.9, below, we have calculated 10, 20 and 50 years ahead the instantaneous rate of change of gas flaring emissions.

Table 2.8 Residual Analysis of CO2 due to Gas Flaring

<table>
<thead>
<tr>
<th>Year</th>
<th>Empirical ROC</th>
<th>DF ROC</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>0.245795</td>
<td>0.017888</td>
<td>0.227907</td>
</tr>
<tr>
<td>1996</td>
<td>-0.04216</td>
<td>0.023983</td>
<td>-0.06614</td>
</tr>
<tr>
<td>1997</td>
<td>-0.05885</td>
<td>0.02968</td>
<td>-0.08853</td>
</tr>
<tr>
<td>1998</td>
<td>-0.59774</td>
<td>0.034904</td>
<td>-0.63264</td>
</tr>
<tr>
<td>1999</td>
<td>0.067929</td>
<td>0.0396</td>
<td>0.028329</td>
</tr>
<tr>
<td>2000</td>
<td>-0.17064</td>
<td>0.043742</td>
<td>-0.21438</td>
</tr>
<tr>
<td>2001</td>
<td>0.050885</td>
<td>0.047322</td>
<td>0.003563</td>
</tr>
<tr>
<td>2002</td>
<td>0.023158</td>
<td>0.050352</td>
<td>-0.02719</td>
</tr>
<tr>
<td>2003</td>
<td>-0.01097</td>
<td>0.052859</td>
<td>-0.06383</td>
</tr>
<tr>
<td>2004</td>
<td>-0.01734</td>
<td>0.054879</td>
<td>-0.07222</td>
</tr>
</tbody>
</table>

Mean of residual -0.09051489
Standard deviation of residual (SD) 0.2209246
Standard error of residual (SE) 0.06986248

Table 2.9 Future Estimation of IROC of CO2 Emission due to Gas Flaring

<table>
<thead>
<tr>
<th>Years</th>
<th>10 years</th>
<th>20 years</th>
<th>50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IROC in future</td>
<td>262.2497</td>
<td>380.1579</td>
<td>733.8823</td>
</tr>
</tbody>
</table>
CO2 in the atmosphere due to cement. For cement the data in metric tons is graphically shown by Figure 2.11 below.

![Figure 2.11 CO2 due to Cement](image)

The differential equation for cement is given by

\[
C(x) + C'(x) = 2.273 \times 10^9 + 3.448 \times 10^6 x + 1.743 \times 10^3 x^2 + 2.937 \times 10^1 x^3. \tag{2.15}
\]

The solution to (2.5.1) is given by

\[
C(x) = 2.277 \times 10^9 + 3.451 \times 10^6 x + 1.744 \times 10^3 x^2 + 2.937 \times 10^1 x^3. \tag{2.16}
\]

A graphical display of the actual data and the solution of the differential equation for cement is given by Figure 2.12.

![Figure 2.12 Model of CO2 due to Cement](image)
The IROC of cement as a function of time is given analytically by

\[ C'(x) = 3.451 \times 10^6 \times 3.488 \times 10^3 x + 8.811 \times 10^1 x^2. \] (2.17)

A graphical display of expression 2.17 is given by Figure 2.13.

The calculated values for the cement model for \( R^2 \) (\( R^2 \) adjusted) and PRESS statistic are given by Table 2.10, below.

![Figure 2.13 IROC of CO2 due to Cement](image)

**Table 2.10 Statistical Evaluation Criteria of CO2 due to Cement**

<table>
<thead>
<tr>
<th>R square</th>
<th>R square adjusted</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8531</td>
<td>0.8426</td>
<td>14076088</td>
</tr>
</tbody>
</table>

The values of \( R^2 \) (\( R^2 \) adjusted) reflect that we have identified a good model along with a PRESS statistic value that is the smallest of several models that we tested. Furthermore, the residual analysis we performed on the proposed differential equation of flaring is given in Table 2.11, below.

As seen from the table, the residuals are extremely small and so is the standard error. These results attest to the good quality of the proposed model for cement.

Thus, in Table 2.12, below, we have calculated 10, 20 and 50 years ahead the instantaneous rate of change of cement emissions.

22
Table 2.11 Residual Analysis of CO2 due to Cement

<table>
<thead>
<tr>
<th>Year</th>
<th>Empirical ROC</th>
<th>DF ROC</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>-0.01297</td>
<td>0.018439</td>
<td>-0.03141</td>
</tr>
<tr>
<td>1996</td>
<td>0.031825</td>
<td>0.02077</td>
<td>0.011055</td>
</tr>
<tr>
<td>1997</td>
<td>0.04258</td>
<td>0.023123</td>
<td>0.019457</td>
</tr>
<tr>
<td>1998</td>
<td>0.01501</td>
<td>0.025473</td>
<td>-0.01046</td>
</tr>
<tr>
<td>1999</td>
<td>0.026395</td>
<td>0.027799</td>
<td>-0.0014</td>
</tr>
<tr>
<td>2000</td>
<td>0.019685</td>
<td>0.030077</td>
<td>-0.01039</td>
</tr>
<tr>
<td>2001</td>
<td>0.010515</td>
<td>0.032289</td>
<td>-0.02177</td>
</tr>
<tr>
<td>2002</td>
<td>0.009024</td>
<td>0.034415</td>
<td>-0.02539</td>
</tr>
<tr>
<td>2003</td>
<td>0.033597</td>
<td>0.03644</td>
<td>-0.00284</td>
</tr>
<tr>
<td>2004</td>
<td>0.049653</td>
<td>0.038351</td>
<td>0.011302</td>
</tr>
</tbody>
</table>

Mean of residual: -0.00618649

<table>
<thead>
<tr>
<th></th>
<th>residual (SD)</th>
<th>residual (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation of residual</td>
<td>0.01692732</td>
<td>0.005352889</td>
</tr>
</tbody>
</table>

Table 2.12 Future Estimation of IROC of CO2 due to Cement

<table>
<thead>
<tr>
<th>Years</th>
<th>10 years</th>
<th>20 years</th>
<th>50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IROC in future</td>
<td>1054.975</td>
<td>1758.297</td>
<td>4925.62</td>
</tr>
</tbody>
</table>

CO2 in the atmosphere due to bunker. For bunker the data in metric tons is graphically shown by Figure 2.14 below.

![Figure 2.14 CO2 due to Bunker](image)

The differential equation for bunker is given by

\[ B(x) + B'(x) = 7.318 \cdot 10^9 \cdot 1.111 \cdot 10^7 \cdot x + 5.597 \cdot 10^3 \cdot x^2 + 9.42 \cdot 10^1 \cdot x^3 \]  \tag{2.1}
The solution to (2.18) is given by

\[ B(x) = 7.329 \times 10^9 + 1.11 \times 10^7 x + 5.6 \times 10^3 x^2 + 9.42 \times 10^1 x^3. \]  

(2.19)

A graphical display of the actual data and the solution of the differential equation for bunker is given by Figure 2.15.

![Figure 2.15 Model of CO2 due to Bunker](image)

The IROC of bunker as a function of time is given analytically by

\[ B'(x) = 1.11 \times 10^7 + 1.12 \times 10^3 x + 2.826 \times 10^2 x^2. \]  

(2.20)

A graphical display of expression 2.20 is given by Figure 2.16.

The calculated values for the solid fuels model for \( R^2 \) (\( R^2 \) adjusted) and PRESS statistic are given by Table 2.13, below.

<table>
<thead>
<tr>
<th>R square</th>
<th>R square adjusted</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8728</td>
<td>0.8637</td>
<td>640380682</td>
</tr>
</tbody>
</table>
The values of $R^2$ (and $R^2_{adj}$) reflect that we have identified a good model along with a PRESS statistic value that is the smallest of several models that we tested. Furthermore, the residual analysis we performed on the proposed differential equation of flaring is given in Table 2.14.

As seen from the table, the residuals are extremely small and so is the standard error. These results attest to the good quality of the proposed model for bunker. Thus, in Table 2.15, below, we have calculated 10, 20 and 50 years ahead the instantaneous rate of change of bunker emissions.
Table 2.15 Future Estimation of IROC of CO2 due to Bunker

<table>
<thead>
<tr>
<th>Years</th>
<th>10 years</th>
<th>20 years</th>
<th>50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IROC in future</td>
<td>-1975.615</td>
<td>-4083.129</td>
<td>-13796.75</td>
</tr>
</tbody>
</table>

CO2 in the atmosphere due to emission. The cumulative emissions of all six attributable variables in thousand metric tons are shown by Figure 2.17 below.

![Cumulative CO2 Emissions from 1950 to 2005](graph.png)

Figure 2.17 Cumulative CO2 Emissions from 1950 to 2005

The differential equation for emission is given by

$$E(t) + E'(t) = 3.73335 \times 10^{10} \times 5.6403 \times 10^{7} \times t + 28405.687 + t^2 \times 4.771 \times t^3. \quad (2.21)$$

The solution to (2.21) is given by

$$E(t) = 3.739 \times 10^{10} \times 5.646 \times 10^{7} \times t + 2.842 \times 10^{4} \times t^2 \times 4.771 \times t^3. \quad (2.22)$$

A graphical display of the actual data and the solution of the differential equation for cumulative emissions are given by Figure 2.18.
Figure 2.18 Model of Cumulative CO2 Emissions
The IROC of emission as a function of time is given analytically by

\[ \frac{d(B(t))}{dt} = 5.906 \times 10^6 \times 5982 \times t \times 1.5051 + t^2 \]  
(2.23)

A graphical display of expression 2.23 is given by Figure 2.19.

Figure 2.19 IROC of Cumulative CO2 Emissions
The calculated values for the solid fuel model for $R^2 \ (R^2 \text{ adjusted})$ $\varepsilon^{-1}$

PRESS statistic are given by Table 2.16, below

27
Table 2.16 Statistical Evaluation Criteria of CO2 Emissions

<table>
<thead>
<tr>
<th>R square</th>
<th>R square adjusted</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9526</td>
<td>0.9498</td>
<td>243305185596</td>
</tr>
</tbody>
</table>

The values of $R^2$ ($R^2$ adjusted) reflect the fact that we have identified a good model along with a PRESS statistic value that is the smallest of several models that we tested. Furthermore, the residual analysis we performed on the proposed differential equation of cumulative CO2 emissions is given in Table 2.17, below.

As seen from the table the residuals are extremely small and so is the standard error. These results attest to the good quality of the proposed model for emission.

Table 2.17 Residual Analysis of CO2 Emissions

<table>
<thead>
<tr>
<th>Year</th>
<th>Empirical ROC</th>
<th>DF ROC</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>0.018860751</td>
<td>0.01035681</td>
<td>0.008503944</td>
</tr>
<tr>
<td>1997</td>
<td>0.023507809</td>
<td>0.01044851</td>
<td>0.013059298</td>
</tr>
<tr>
<td>1998</td>
<td>-0.008980365</td>
<td>0.01055562</td>
<td>-0.019535981</td>
</tr>
<tr>
<td>1999</td>
<td>0.016188277</td>
<td>0.01067738</td>
<td>0.005510893</td>
</tr>
<tr>
<td>2000</td>
<td>0.033282206</td>
<td>0.01081308</td>
<td>0.022469128</td>
</tr>
<tr>
<td>2001</td>
<td>-0.015679749</td>
<td>0.01096196</td>
<td>-0.026641713</td>
</tr>
<tr>
<td>2002</td>
<td>0.012155809</td>
<td>0.01112331</td>
<td>0.001032500</td>
</tr>
<tr>
<td>2003</td>
<td>-0.004415916</td>
<td>0.01129638</td>
<td>-0.015712294</td>
</tr>
<tr>
<td>2004</td>
<td>0.019475149</td>
<td>0.01148044</td>
<td>0.007994709</td>
</tr>
<tr>
<td>2005</td>
<td>0.009044926</td>
<td>0.01167476</td>
<td>-0.002629837</td>
</tr>
</tbody>
</table>

Mean of residual -0.0005949354

Standard deviation of residual (SD) 0.1558048

Standard error of residual (SE) 0.004926982

Thus, in Table 2.18, below, we have calculated 10, 20 and 50 years ahead the instantaneous rate of change of emission.

Table 2.18 Future Estimation of IROC of Cumulative CO2 Emissions

<table>
<thead>
<tr>
<th>Years</th>
<th>10 years</th>
<th>20 years</th>
<th>50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IROC in future</td>
<td>41400.43</td>
<td>51245.62</td>
<td>97956.83</td>
</tr>
</tbody>
</table>

The cumulative CO2 in atmosphere. The cumulative CO2 in atmosphere is shown by Figure 2.20 below.
The differential equation for emission is given by

\[ A(x) + A'(x) = 1.953 \quad 10^7 \quad 2.933 \quad 10^4 \quad x \quad 1.468 \quad 10^1 \quad x^2 \quad 2.448 \quad 10^3 \quad x^3 \]  

(2.21)

The solution to (2.21) is given by

\[ A(x) = 1.956 \quad 10^7 \quad 2.936 \quad 10^4 \quad x \quad 1.469 \quad 10^1 \quad x^2 \quad 2.448 \quad 10^3 \quad x^3 \]  

(2.22)

A graphical display of the actual data and the solution of the differential equation for cumulative emissions is given by Figure 2.21.

The IROC of emission as a function of time is given analytically by

\[ A'(x) = 2.936 \quad 10^4 \quad + \quad 2.938 \quad 10^1 \quad x \quad 7.343 \quad 10^3 \quad x^2 \]  

(2.:}
A graphical display of expression 2.23 is given by Figure 2.22.

![Graph of IROC of Cumulative CO2 in the Atmospheres](image)

Figure 2.22 IROC of Cumulative CO2 in the Atmospheres

The calculated values for the solid fuel model for \( R^2 \) (\( R^2 \) adjusted) and PRESS statistic are given by Table 2.19, below.

<table>
<thead>
<tr>
<th>R square</th>
<th>R square adjusted</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9992</td>
<td>0.9991</td>
<td>2289.919</td>
</tr>
</tbody>
</table>

The values of \( R^2 \) (\( R^2 \) adjusted) reflect the fact that we have identified a good model along with a PRESS statistic value that is the smallest of several models that we tested. Furthermore, the residual analysis we performed on the proposed differential equation of cumulative CO2 in the atmospheres is given in Table 2.20, below.

As seen from the table, the residuals are extremely small and so is the standard error. These results attest to the good quality of the proposed model for emission.
Table 2.20 Residual Analysis of CO2 in the Atmospheres

<table>
<thead>
<tr>
<th>Year</th>
<th>Empirical ROC</th>
<th>DF ROC</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>0.005573</td>
<td>0.004568</td>
<td>0.001005</td>
</tr>
<tr>
<td>1996</td>
<td>0.004877</td>
<td>0.004565</td>
<td>0.000312</td>
</tr>
<tr>
<td>1997</td>
<td>0.003088</td>
<td>0.004559</td>
<td>-0.00147</td>
</tr>
<tr>
<td>1998</td>
<td>0.00789</td>
<td>0.004555</td>
<td>0.00334</td>
</tr>
<tr>
<td>1999</td>
<td>0.004582</td>
<td>0.004537</td>
<td>4.55E-05</td>
</tr>
<tr>
<td>2000</td>
<td>0.003177</td>
<td>0.004521</td>
<td>-0.00134</td>
</tr>
<tr>
<td>2001</td>
<td>0.004168</td>
<td>0.004502</td>
<td>-0.00033</td>
</tr>
<tr>
<td>2002</td>
<td>0.005606</td>
<td>0.004479</td>
<td>0.001127</td>
</tr>
<tr>
<td>2003</td>
<td>0.006808</td>
<td>0.004454</td>
<td>0.002354</td>
</tr>
<tr>
<td>2004</td>
<td>0.004632</td>
<td>0.004425</td>
<td>0.000207</td>
</tr>
<tr>
<td></td>
<td>Mean of residual</td>
<td>0.0005242005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Standard deviation of residual (SD)</td>
<td>0.001507517</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Standard error of residual (SE)</td>
<td>0.0004767186</td>
<td></td>
</tr>
</tbody>
</table>

Thus, in Table 2.21, below, we have calculated 10, 20 and 50 years ahead the instantaneous rate of change of emission.

Table 2.21 Future Estimation of IROC of Cumulative CO2 in the Atmospheres

<table>
<thead>
<tr>
<th>Years</th>
<th>10 years</th>
<th>20 years</th>
<th>50 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>IROC in future</td>
<td>18.60843</td>
<td>15.86782</td>
<td>-1.165586</td>
</tr>
</tbody>
</table>

2.4 Conclusions

In the present study we have developed the actual differential equations that characterize the behavior of each of the six attributable variables that constitute the carbon dioxide emissions, namely, gas fuels, liquid fuels, solid fuels, flaring, cement, and bunker. We have developed the differential system of the sum of the six attributable variables that constitute CO2 emissions using actual data from 1950 to 2005 for continental United States. In addition to have given the analytical form for each variable, we have used three different statistical procedures, namely \( R^2 \) (\( R^2 \) adjusted), PRESS statistic and residual analysis to evaluate the quality of the proposed differential methods. All these statistical procedures attest to the quality of the proposed differential systems. Finally, we have used these models to make 10,
20, and 50 year prediction to the rate of change of the entities that constitute CO2 emissions. This information should be useful for strategic planning and formulating policies to assist in the problem of GLOBAL WARMING.

In the present study we have developed the actual differential equations that characterize the behavior of each of the five attributable variables that constitute the carbon dioxide in the atmosphere, namely, gas fuels, liquid fuels, solid fuels, flaring, cement, and bunker. We have developed the differential system of the total CO2 in the atmospheres using actual data from 1959 to 2004 for the continental United States. In addition to having given the analytical form for each variable, we have used three different statistical procedures, namely $R^2$, $R^2$ adjusted, PRESS statistic and residual analysis to evaluate the quality of the proposed differential methods. All these statistical procedures attest to the quality of the proposed differential systems. Finally, we have used these models to make 10, 20, and 50 year predictions to the rate of change of the entities that constitute CO2 in the atmosphere. This information should be useful for strategic planning and formulating policies to assist in the problem of GLOBAL WARMING.
Chapter 3:

Attributable variables with interactions that contribute to carbon dioxide in the atmosphere

3.1 Background and Data

The proposed model that we are developing takes into consideration individual contributions and interactions along with higher order contributions if applicable. In defining the analytical structure of each variable, we used real yearly data that have been collected from 1959 to 2004 for the continental United States.

Figure 3.1 The Attributable Variables of CO2 in the Atmosphere

We proposed an overall model to modeling the CO2 in the atmosphere and all possible attributable variables. Those individual variables with the significant
interactions are ranked according to their contributions to the CO2 in the atmosphere and are listed in Figure 3.1, above.

The proposed statistical model is useful in predicting the CO2 in the atmosphere given the information of attributable variables. It has been statistically evaluated using R square, R square adjusted, PRESS statistic and residual analysis. Finally, its usefulness has been illustrated by utilizing different combinations of various attributable variables. To our knowledge, no such model has been developed under the proposed analytical structure. In addition, we rank the attributable variables according to their CO2 contributions in the atmosphere.

The illustration of Carbon dioxide circulation process in the atmosphere that was developed by scientists at the Oak Ridge National laboratory is given by Figure 3.2, below. (from ICPP’s report)
There are other possible attributable variables namely, deforestation and destruction of biomass and soil carbon (D), terrestrial plant respiration (R), respiration from soils and decomposers (S), the flux from oceans to atmosphere (O), terrestrial photosynthesis (P), the flux from atmosphere to oceans (I), the burial of organic carbon and limestone carbon in sediments and soils (B). Due to our data base limitation, we will not consider them in our current model. We will update the model once we have access to those data.

3.2 Methodology

We proceed to develop a statistical model taking into consideration the eight attributable variables as presented previously. The form of the statistical model is given by CO2 in the atmosphere as a function of temperature. Thus, the statistical form of the model with all possible interactions will be

\[ \text{CO}_2 = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \ldots + \beta_i A_i + \beta_j B_j + \ldots \]  

Here the \( \beta \) and \( A \) are the coefficients and \( A \) are the first order term of the attributable variables and \( B \) are the possible interactions and higher order terms. The object is to develop the most representative estimate of the above model based on available data. In the present study we will focus on using atmospheric CO2 as response and only six attributable variables as our independent variables.

The data comes from Oak Ridge National Lab: Division of U.S. Department of Energy. The plot of CO2 in the atmosphere is shown in Figure 3.3, below. The air samples collected at Mauna Loa Observatory, Hawaii and the data unit is in ppmv.

One of the underlying assumptions to construct the above model 3.1 is that the response variable should follow Gaussian distribution. We know the CO2 in the atmosphere are not follow Gaussian distribution which can be clearly seen from the QQ plot shown by Figure 3.4, below.
We will utilize Box-Cox transformation to the CO2 atmosphere data to filter the data to be normally distributed. After we proceed with the Box-Cox transformation, the results are shown in Table 3.1, below.

![Graph](image_url)

**Figure 3.3 Yearly CO2 in Atmosphere Data at Mauna Loa**

After the Box-Cox filter, we retest the data and it shows our data will follow normal distribution; thus, we proceed to estimate the coefficients of the contributable variables for the transformed CO2 atmosphere data in the equation 3.1.

<table>
<thead>
<tr>
<th>Est. Power</th>
<th>Std. Err</th>
<th>Wald (Power=0)</th>
<th>Wald (Power=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.3763</td>
<td>.9609</td>
<td>-2.4729</td>
<td>-3.5136</td>
</tr>
</tbody>
</table>

We can proceed to estimate the approximate coefficients of the contributable variables for transformed CO2 in the atmosphere and obtain the coefficient of all
possible interactions. At the same time, we can determine the significant contributions of both attributable variables and interactions.

![Normal Q-Q Plot](image)

**Figure 3.4 QQ Plot for Testing Normality**

We begin with six attributable variables as previously defined, such as Ga, So, Li, Fl, Ce, and Bu and fifteen 2nd order interactions between each pair. To develop the models, initially we start building our model with 21 total terms that include initial contribution of attributable variables and all possible interactions. We construct twenty-four such models.

During our statistical analysis in the estimation process, we found only one out of six attributable variables significantly contribute and five interaction terms. Thus the result of estimation of equation 3.1 is given by equation 3.2 as follows

\[
\hat{C}_O^2 = 3.196 \times 10^{-9} - 2.586 \times 10^{14} \text{Ga} - 1.296 \times 10^{15} \text{Li} - 1.939 \times 10^{-14} \text{FL} \\
+ 6.922 \times 10^{14} \text{Ce} - 8.961 \times 10^{15} \text{Bu} - 2.107 \times 10^{19} \text{Ga FL} + 5.593 \times 10^{19} \text{Li FL} \\
- 2.559 \times 10^{19} \text{Li Ce} - 5.822 \times 10^{18} \text{FL Bu} + 2.049 \times 10^{18} \text{Ce Bu}
\]
We will utilize the initial transformation that we used to transform the response data to get the result in equation 3.3 by taking the \((-0.4208)\)'s power on both sides of the equation 3.2.

\[

cO_2 = (3.196 \times 10^{-9} - 2.586 \times 10^{-17} \text{Ga} - 1.296 \times 10^{-15} \text{Li} - 1.939 \times 10^{-14} \text{FL} \\
+ 6.922 \times 10^{-14} \text{Ce} - 8.961 \times 10^{-15} \text{Bu} - 2.107 \times 10^{-19} \text{Ga} \text{ FL} + 5.593 \times 10^{-19} \text{Li} \text{ FL} \\
- 2.559 \times 10^{-19} \text{Li} \text{ Ce} - 5.822 \times 10^{-18} \text{FL} \text{ Bu} + 2.049 \times 10^{-18} \text{Ce} \text{ Bu})^{0.4208}
\] (3.3)

This proposed nonlinear statistical model identifies the following attributable variables. We can find Ca, Li, FL, Ce and Bu significantly contribute to the CO2. Furthermore, we have identified the following interactions that have been show statistically contribute to CO2 namely Ga*FL, Li*FL, Li*Ce, FL*Bu and Ce*Bu. We summarized our model in the Figure 3.5.

The proposed underlying statistical model is high in quality. It has been evidenced by high value of both R square and R square adjusted which are the key criteria to evaluate the model fitting. The regression sum of squares (SSR), also called the explained sum of squares, is the variation that is explained by the regression model. The sum of squared errors (SSE), also called the residual sum of squares, is the variation that is left unexplained. The total sum of squares (SST) is proportional to the sample variance and equals the sum of SSR and SSE. The coefficient of determination $R^2$ is defined as the proportion of the total response variation that is explained by the model. It provides an overall measure of how well the model fits. R-square is SSR/SST. R-square adjusted will adjust for degree of freedom of the model and it works better when we have many parameters. R-square adjusted is $R_{adj}^2 = 1 - \frac{SSE}{DF(SSE)} \frac{SST}{DF(SST)}$. 

38
The prediction of residual error sum of squares (PRESS) statistics will evaluate how good the estimation will be if each time we remove one data point and

\[ PRESS = \left( \frac{y \cdot \hat{y}}{H_{n}} \right)^{2}, \]  


Table 3.2 PRESS Statistics for Best Three Models

<table>
<thead>
<tr>
<th>Model number</th>
<th>PRESS value</th>
<th>Rank of the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>3.414703e-20</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>3.523170e-20</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>8.1202e-20</td>
<td>3</td>
</tr>
</tbody>
</table>

For our final model the R squared is 0.9963 and R squared adjusted is 0.9953. Both R squared and R squared adjusted are very high (more than 90%) and these two are very close to each other. This shows our model’s R squared increase in not due to the increase of the parameters estimates but the good quality of the proposed model to predict CO2 in the atmosphere given values of the identified attributable variables. Secondly, the PRESS statistics results support the fact that the proposed model is of high quality. We will list the best three models' PRESS statistic out of total 28 and the result is in Table 3.2. From the table it is clear that the best model is number 28, which is our final model.

Furthermore, R square and R square adjusted are calculated for those 28 models which are of interest but the proposed model gives the best possible estimates of the CO2 in the atmosphere. We just present the best possible model’s statistical evaluation criteria in Table 3.3.

Table 3.3 Statistical Evaluation Criteria

<table>
<thead>
<tr>
<th>R square</th>
<th>R square adjusted</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9963</td>
<td>0.9953</td>
<td>3.414703e-20</td>
</tr>
</tbody>
</table>
Figure 3.5 CO2 in the Atmosphere Attributable Variable Diagram
Figure 3.6 CO2 in the Atmosphere Attributable Variable Contribution Diagram

Table 3.4 Rank of Variable According to Contributions

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Liquid</td>
</tr>
<tr>
<td>2</td>
<td>Liquid: Cement</td>
</tr>
<tr>
<td>3</td>
<td>Cement: Bunker</td>
</tr>
<tr>
<td>4</td>
<td>Bunker</td>
</tr>
<tr>
<td>5</td>
<td>Cement</td>
</tr>
<tr>
<td>6</td>
<td>Gas Flares</td>
</tr>
<tr>
<td>7</td>
<td>Gas Fuels</td>
</tr>
<tr>
<td>8</td>
<td>Gas Fuels: Gas Flares</td>
</tr>
<tr>
<td>9</td>
<td>Liquid: Gas Flares</td>
</tr>
<tr>
<td>10</td>
<td>Gas Flares: Bunker</td>
</tr>
</tbody>
</table>

The Table 3.4 ranks the attributable variables with respect to their contribution to CO2 in the atmosphere. As we expected, Li ranks number one which is one of the attributable variables from the emissions from fossil fuels. The individual contributions with interactions are shown in Figure 3.6. We ranked those terms by their percentage of contribution to CO2 in the atmosphere.
3.3 VALIDATION OF THE PROPOSED MODEL

We will utilize two methods to do the model validation. The first method is to use the proposed model to calculate the predicted value for each individual data and then calculate the residuals. The residual is defined as the original value minus the predicted value. Table 3.5 shows the last ten residuals out of the total one hundred fifty-five residuals.

Table 3.6 shows the mean of the residuals is -.0286, variance of the residuals is 1.588, standard deviation is 1.26 and standard error of the residuals is .1012.

<table>
<thead>
<tr>
<th>No</th>
<th>Residual Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>-1.142226e-12</td>
</tr>
<tr>
<td>38</td>
<td>-4.122377e-12</td>
</tr>
<tr>
<td>39</td>
<td>1.764721e-11</td>
</tr>
<tr>
<td>40</td>
<td>4.939785e-12</td>
</tr>
<tr>
<td>41</td>
<td>-3.126147e-11</td>
</tr>
<tr>
<td>42</td>
<td>2.243955e-11</td>
</tr>
<tr>
<td>43</td>
<td>3.234128e-11</td>
</tr>
<tr>
<td>44</td>
<td>1.167639e-11</td>
</tr>
<tr>
<td>45</td>
<td>-2.728478e-11</td>
</tr>
<tr>
<td>46</td>
<td>-1.357971e-11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean of Residual</th>
<th>3.645909 e-28</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Deviation of Residual (SD)</td>
<td>2.020493e-11</td>
</tr>
<tr>
<td>Standard Error of Residual (SE)</td>
<td>5.832662e-12</td>
</tr>
</tbody>
</table>

The second method we will utilize is the cross validation. The basic idea is we will save some part of the data as validation part. We construct our model using only the data left and the constructed model will be same structure as our proposed model with only coefficients being different. We will test the quality of model using three settings.
We will first randomly divide the data into two data sets of same size. Then we will use one to construct the model and then use this model to predict the value using other data set’s attributable variables. Then we will switch the two data sets and repeat the procedure. The mean of all residuals is 1.052026e-21.

Second we will divide the data set into six small data sets and use five of them to construct the model and validate the model using the sixth one. Then we will repeat the same procedure for each of the six small data sets. The mean of all residuals is 8.378600e-22.

Thirdly, we will divide the data set into 46 data sets and use all 45 sets to construct the model and validate the model using the one left out. Then we repeat the procedure 46 times. Table 3.7 shows the last ten residuals out of the total 46 residuals.

The mean of the residuals is 7.423267e-22, variance of the residual is 7.007e-43, standard deviation is 8.371e-22 and standard error of the residuals is 8.370868e-22.

<table>
<thead>
<tr>
<th>No</th>
<th>Residual Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>3.445021e-24</td>
</tr>
<tr>
<td>38</td>
<td>3.113447e-23</td>
</tr>
<tr>
<td>39</td>
<td>7.893401e-22</td>
</tr>
<tr>
<td>40</td>
<td>3.354917e-23</td>
</tr>
<tr>
<td>41</td>
<td>2.003004e-21</td>
</tr>
<tr>
<td>42</td>
<td>1.836528e-21</td>
</tr>
<tr>
<td>43</td>
<td>1.760729e-21</td>
</tr>
<tr>
<td>44</td>
<td>1.938752e-22</td>
</tr>
<tr>
<td>45</td>
<td>1.886689e-21</td>
</tr>
<tr>
<td>46</td>
<td>5.623131e-22</td>
</tr>
</tbody>
</table>

**Table 3.7 Residual Analysis for Cross Validation**

**3.4 CONCLUSIONS**

In the present study, we have performed parametric analysis for CO2 in the atmosphere. The initial measurement of CO2 in the atmosphere was collected at Mauna Loa Observatory, Hawaii (C.D. Keeling, T.P. Whorf, 2005). Those data do r
follow normal probability distribution. Thus, we transform the response of the data by
using Box-Cox transformation that resulted in make the CO2 being normal. We
proceed to develop a “nonlinear” statistical model (nonlinear in terms of the higher
power of the response variable). Through the process of developing the statistical
model, we have found that only five variables, namely, Liquid, Bunker, Cement, Gas
flares and Gas fuels significantly contribute to the CO2 in the atmosphere with five
interactions among them. The proposed statistical model was evaluated using the R-
square, R-square adjusted and PRESS statistics. We also provide model cross
validation by hiding some part of the data and estimate them from the rest of the data.
The mean residuals of those cross validation are extremely small. All criteria results
support the high quality of the developed statistical model. Furthermore we have
ranked the individual attributable variable and interaction according to their
contribution to CO2 in the atmosphere.

This model can be used to obtain a good estimate of CO2 in the atmosphere
knowing only the five significantly attributable variables mentioned above.

We can conclude from our extensive statistical analysis that there are only
five significant attributable variables to the CO2 in the atmosphere namely, Gas fuels,
Gas flares, Bunker, Liquid and Cement. Furthermore, we also tested all possible 2nd
order interactions of all attributable variables and we found only five interactions that
significantly contribute to CO2 in the atmosphere, namely, Liquid with Cement,
Cement with Bunker, Gas Fuels with Gas Flares, Liquid with Gas Flares and Gas
Flares with Bunker. Thus, one may obtain a good estimate of the CO2 in the
atmosphere by knowing the measurement of Cement and those five interactions.

One can utilize the above model equation 3.2 to perform surface response
analysis to identify the values of the contributable variables that will minimize the
CO2 in the atmosphere.
Chapter 4

Probabilistic comparison of survival analysis models using simulation and cancer data

4.1 Background and Data

Wikipedia defines survival analysis as a branch of statistics which deals with death in biological organisms and failure in mechanical systems. Scientists have developed and used many different probabilistic survival analysis methods including parametric, nonparametric and semi-parametric models. In the present study we will compare all commonly used methods and propose which ones give the best probabilistic survival results.

The first part of our study is based on simulating data from a well defined probability failure distribution by identifying the sample size so that the maximum likelihood estimates converge to the assumed parametric values in a Monte Carlo simulation procedure. Using this information we develop and compare the parametric estimated probabilistic survival function with the Kernel density (nonparametric), and the popular Kaplan-Meier (KM) model.

The second part of our study uses actual survival time of breast cancer data to compare the above mentioned survival models, in addition to the Cox Proportional (Cox PH) survival hazard function.

Upon completing the evaluation, we will propose a ranking of the analytical methods evaluated for performing survival analysis. The breast cancer data that we used was given by N. A Ibrahim where the analysis and results were published (Ibrahim et. al., 2008).
4.2 Literature Review

Survival analysis is very useful for cancer research. Many researchers have contributed to this subject. Kaplan Meier empirical type of survival model is first constructed in 1958 (Kaplan and Meier, 1958). The Mantel Haenszel test for survival analysis is proposed in 1959 (Mantel and Haenszel, 1959). The generalized Wilcoxon test is developed in 1965 (Gehan, 1965) and this test is more powerful than the Cox proportional hazard’s test when the proportional hazard assumption is violated early on. The Cox proportional hazards (PH) model for survival data is introduced in 1972 (Cox, 1972). A class of rank test procedures for censored survival data is presented in 1982 (Harrington and Fleming, 1982). Therneau et. al. 1990 and 2000 studied the Cox model and residuals for survival models. The regression with frailty in survival analysis is discussed in 1991 (McGilchrist and Aisbett, 1991, McGilchrist, 1993). A semiparametric estimation of random effects using the Cox model is provided in 1992 (Klein, 1992). The accelerated life testing model is studied extensively in 2000 (Qiu and Tsokos, 2000). A semi-parametric accelerated failure model is introduced in 2002 (Shang and Jeremy 2002). An analytical approach on cure rate model based on uncensored data is discussed in 2006 (Uddin et al., 2006). Using decision tree for competing risks for breast cancer is discussed in 2008 (Ibrahim et al., 2008). There are several researches that had been done to determine the factors that are contribute to the relapse time of the breast cancer (Eleni and Gabriel, 2008; Habibi et al., 2008; Freedman et al., 2009; Brawley, 2009).

4.3 Methodology

4.3.1 Survival Analysis using Simulation

For our parametric Monte Carlo simulation, we shall assume that the failure data is being probabilistically characterized by the two parameter gamma probability density function (pdf), given by
\[ f(x, \theta, \alpha) = \frac{x^\alpha e^{-x}}{\Gamma(\alpha)}, \quad \alpha > 0, \quad x > 0 \] 

(4.1)

where \( \alpha \) and \( \theta \) are the shape and scale parameter, respectively and

\[ ( \alpha ) = \int_0^\alpha t^{\alpha - 1} e^{-t} dt. \]

The cumulative distribution function (CDF) of 2.1 is given by

\[ F(x, \theta, \alpha) = \frac{x^\alpha \Gamma(\alpha)}{\Gamma(k)}, \quad x > 0, \quad \alpha > 0 \] 

(4.2)

where \( \Gamma(\alpha, x) = \int_0^x t^{\alpha - 1} e^{-t} dt \) is the lower incomplete gamma function.

The survival function of the gamma pdf is given by

\[ S(x, \theta, \alpha) = 1 - \frac{x^\alpha \Gamma(\alpha)}{\Gamma(k)}, \quad x > 0, \quad \alpha > 0, \quad \alpha > 0 \] 

(4.3)

and the hazard function is of the form

\[ h(x) = \frac{-e^{-x}x^{\alpha - 1}}{\Gamma(\alpha)}, \quad x > 0, \quad \alpha > 0. \] 

(4.4)

Also, the cumulative hazard function is given by

\[ H(x) = \log(1 - \frac{x^\alpha \Gamma(\alpha)}{\Gamma(k)}), \quad \alpha > 0, \quad \alpha > 0. \] 

(4.5)

where \( \Gamma(\alpha, x) \) is the incomplete gamma function.

For the first part of our study we assume that parameter \( \alpha = 3.0, \)

\( = 2.0 \) and rate parameter \( \theta = 0.5 \). We simulated a sample of \( n=300 \) failure time

where the maximum likelihood estimates resulted in \( \hat{\alpha} = 2.984, \hat{\theta} = 2.009 \) and

\( \hat{\alpha} = \left( \hat{\theta} \right)^{-1} = 0.498 \) which closely converge to the assumed true parameters. Thus, \( \hat{f} \)
n=300 failures, we have a very good random sample to begin parametrically our evaluation process.

Thus, the parametric true survival and hazard functions are given by

\[ S(t,3,2) = 1 - t^2 \frac{e^{-\frac{t}{3}}}{(3)^2}, \]  

(4.6)

and the true hazard function is given by

\[ h(t) = \frac{t^2 e^{-\frac{t}{3}}}{(3)^2}, \quad t > 0. \]  

(4.7)

The estimated parametric survival function is given by

\[ \hat{S}(t, \hat{\eta}, \hat{\mu}) = 1 - t^2 \frac{e^{-\frac{t}{2.0093}}}{(2.984)^2}, \]  

(4.8)

and the estimated parametric hazard function is given by

\[ \hat{h}(t) = \frac{t^2 e^{-\frac{t}{2.0093}}}{(2.984)^2}, \]  

(4.9)

respectively.

For comparison purposes, we shall refer to the survival functions, \( S(t, 3, 2) \), as the true parametric probabilistic survival curve and \( \hat{S}(t; 2.984, \hat{\eta} = 0.4977) \), as the parametric estimates. Figure 4.1, below, gives a graphical display of the two probabilistic survival curves and Figure 4.2, below, gives the corresponding estimate of the hazard function, \( \hat{H}(t) \).

It is clear that both estimated plots and true parametric plots are almost identical as a function of age.

Probability residual analysis

In order to clearly show the definition of the probability residuals, we made the Figure 4.3 below from a random gamma distribution. We digitized the time into n=3^-
single point. At each single point we define the difference between the two survival curves as our probability residuals. $r_i = \hat{S}_{true}(t_i) - \hat{S}_{fitted}(t_i)$ for $i=1,2,3,\ldots,n.$

![Figure 4.1 Survival Plots for True and Fitted Parametric Analysis](image1.png)

Figure 4.1 Survival Plots for True and Fitted Parametric Analysis

![Figure 4.2 Cumulative Hazard Plots for True and Fitted Parametric Analysis](image2.png)

Figure 4.2 Cumulative Hazard Plots for True and Fitted Parametric Analysis
Figure 4.3 Illustration of The Definition of Probability Residuals

Then we proceed to calculate the mean probability residual, the sample variance, the sample standard deviation, and sample standard error. Thus, the mean of the probability residual is 0.000655, sample variance is 2.675e-07, sample standard deviation is 0.00052 and standard error is 2.986e-05. These numbers attest to the quality of the suggest model.

Kaplan-Meier Method

The Kaplan-Meier method is the most popular in developing the survival functions for a given set of failure times. The survival function, S(t), is the probability that an item from a given population will have a survival time exceeding t. Let us consider a random sample of size n of the failure observed times until death, that is, \( t_1, t_2, \ldots, t_n \) and arranging them in the following manner

\[
t_1 \quad t_2 \quad t_3 \quad \ldots \quad t_{n-1} \quad t_n.
\]

Define \( n_i \) as the number of patients at risk just prior to time \( t_i \) and let \( d_i \) be the number of deaths at exactly time \( t_i \).

The estimate of the survival function of the Kaplan-Meier model is given by
\[ \hat{S}(t) = \prod_{t_i < t} \frac{ni - di}{ni}, \] (4.10)

The estimate of the cumulative hazard function is given by
\[ \hat{H}(t) = -\ln\left( \prod_{t_i < t} \frac{ni - di}{ni} \right). \] (4.11)

For the Monte Carlo simulation of \( n=300 \) failure times we have the plots in Figure 4.4, the true parametric form of \( S(t) \) and with the Kaplan-Meier estimate of the survival curve along with the 95% confidence limits.

Figure 4.5, below, displays the estimated hazard function of the KM model along with the true parametric plot with 95% confidence limits.

Probability Residual Analysis
Consider the difference between true parametric survival curve and the estimate \( S(t) \) as our probability residual as defined before. Thus, the mean of the probability residual is -0.00345, sample variance is 0.000267, sample standard deviation is 0.01635304 and standard error is 0.000944143. Clearly as expected the parametric survival models give continuously better results than the Kaplan-Meier model.

![Figure 4.4 The True and KM Survival Curve with 95% C.I.](image-url)
Figure 4.5 The Cumulative Hazard Plot for True and KM Method with 95% C.I.

Kernel Density Estimation

A very powerful nonparametric method that has been used extensively to estimate the probability density of a certain data that does not follow any well known classical pdf is the Kernel Density estimation.

For a given set of data $x_1, x_2, \ldots, x_n$, an independent and identically distributed sample of a random variable, then the kernel density estimate of the probability density function is given by

$$
\hat{f}_h(x) = \frac{1}{nh} \sum_{j=1}^{N} K\left(\frac{x - x_j}{h}\right),
$$

(4.12)

and $K$ is the kernel and $h$ is the optimal bandwidth.

Given below, Table 4.1 is a list of the most commonly used kernels.
### Table 4.1 The Most Used Kernel Densities

<table>
<thead>
<tr>
<th>Name of kernel</th>
<th>Math form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>( K(u) = \frac{1}{2} \mathbb{I}_{</td>
</tr>
<tr>
<td>Triangle</td>
<td>( K(u) = (1 -</td>
</tr>
<tr>
<td>Epanechnikov</td>
<td>( K(u) = \frac{3}{4} (1 - u^2) \mathbb{I}_{</td>
</tr>
<tr>
<td>Quartic</td>
<td>( K(u) = \frac{15}{16} (1 - u^2)^2 \mathbb{I}_{</td>
</tr>
<tr>
<td>Triweight</td>
<td>( K(u) = \frac{35}{32} (1 - u^2)^3 \mathbb{I}_{</td>
</tr>
<tr>
<td>Gaussian</td>
<td>( K(u) = \frac{1}{\sqrt{2}} e^{-\frac{1}{2} u^2} \mathbb{I}_{</td>
</tr>
<tr>
<td>Cosine</td>
<td>( K(u) = \frac{1}{4} \cos \left( \frac{u}{2} \right) \mathbb{I}_{</td>
</tr>
</tbody>
</table>

The most frequently used optimal bandwidth is shown in Table 4.2, below.

### Table 4.2 The Most Used Optimal Bandwidth

<table>
<thead>
<tr>
<th>Name of optimal bandwidth</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal 0 (Nrd0)</td>
<td>Represents the bandwidth of a Gaussian kernel. The numerical value is 0.9<em>min[standard deviation, Interquartilerange (IQR)] / (1.34</em>(sample size)^-1/5))</td>
</tr>
<tr>
<td>Normal (nrd)</td>
<td>The numerical value is 1.06<em>min(SD, IQR) / (1.34</em>(sample size)^-1/5))</td>
</tr>
<tr>
<td>unbiased cross-validation (ucv)</td>
<td>For unbiased cross-validation</td>
</tr>
<tr>
<td>biased cross-validation (bcv)</td>
<td>For biased cross-validation</td>
</tr>
<tr>
<td>Select (SJ)</td>
<td>select the bandwidth using pilot estimation of derivatives</td>
</tr>
</tbody>
</table>
Analytical Form

We run the models for all the combinations of different kernels and optimal bandwidth. The Epanechnikov kernel and the proposed optimal bandwidth gave the best results.

The Kernel Density survival function is given by

\[
S_h(x) = 1 \frac{1}{N} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right),
\]

where \(K\) is the kernel and \(h\) is the optimal bandwidth, respectively.

The hazard function of Kernel Density \(S_h(x)\) method is given by

\[
h_h(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{K\left(\frac{x - x_i}{h}\right)}{1 \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)},
\]

with

\[
h = \frac{1.06 \min\{SD(x), IQR(x)\}}{1.34n^{\frac{1}{5}}},
\]

Thus the estimated value of \(S_h(x)\) is given by

\[
\hat{S}_h(x) = 1 \frac{1}{N} \frac{1.06 \min\{SD(x), IQR(x)\}}{1.34n^{\frac{1}{5}}} \sum_{i=1}^{N} \frac{3}{4} \left(1 - \left(\frac{x - x_i}{1.06 \min\{SD(x), IQR(x)\}}\right)^2\right),
\]

and the estimate of the hazard function is given by

\[
\hat{h}_h(x) = \frac{1}{N} \frac{1.06 \min\{SD(x), IQR(x)\}}{1.34n^{\frac{1}{5}}} \sum_{i=1}^{N} \frac{3}{4} \left(1 - \left(\frac{x - x_i}{1.06 \min\{SD(x), IQR(x)\}}\right)^2\right),
\]

54
Thus, the estimate of the survival function, \( \hat{S}_h(x) \) for the KD method is given by Figure 4.6, below. It is clear that the KD approach to \( \hat{S}(t) \) is almost identical to the parametric survival function.

Thus, it is clear from the graph, Figure 4.7, that the true \( \hat{H}(t) \) and the Kernel density estimates are approximately the same.

Probability Residual Analysis

Consider the difference between the true and parametric survival curve as our probability residual. Then the sample mean of the probability residual is -0.0023, sample variance is 0.0001, sample standard deviation is 0.0163 and standard error is 0.000579.

Table 4.3 below gives us the summary and comparison of the three probability residuals.

![Figure 4.6 The Survival Curves for The True Parametric and KD Model](image-url)
**Figure 4.7 The Cumulative Hazard Plot for True and KD Method**

**Table 4.3 Residual Analysis**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>SD</th>
<th>SE</th>
<th>Rank of Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted parametric vs.</td>
<td>0.00065</td>
<td>2.67e-07</td>
<td>0.00052</td>
<td>2.986e-05</td>
<td>1</td>
</tr>
<tr>
<td>True parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaplan-Meier vs. True</td>
<td>-0.0034</td>
<td>0.00027</td>
<td>0.0163</td>
<td>0.000944</td>
<td>3</td>
</tr>
<tr>
<td>parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel density vs.</td>
<td>-0.0023</td>
<td>0.0001</td>
<td>0.01</td>
<td>0.000579</td>
<td>2</td>
</tr>
<tr>
<td>True parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4.4 Residual Analysis Repeat 1000 times**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>SD</th>
<th>SE</th>
<th>Rank of Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted parametric vs.</td>
<td>-0.00621</td>
<td>4.61e-05</td>
<td>0.006787</td>
<td>0.0003919</td>
<td>1</td>
</tr>
<tr>
<td>True parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaplan-Meier vs. True</td>
<td>-0.00143</td>
<td>0.000269</td>
<td>0.015754</td>
<td>0.0009095</td>
<td>3</td>
</tr>
<tr>
<td>parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel density vs.</td>
<td>-0.00025</td>
<td>0.000297</td>
<td>0.016040</td>
<td>0.0009260</td>
<td>2</td>
</tr>
<tr>
<td>True parametric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
From the above Table 4.3 and Table 4.4 we can see that the parametric model as expected is much better than other two. KM and KD are very close using the mean of the probability residuals but KD’s standard deviation and standard error is about 50% smaller than the popular KM model. Therefore, if parametric analysis is not justified, we recommend that the KD should be used for developing the survival model.

From this Figure 4.8 we can observe that KD is much smoother than KM and is closer to true in the middle but in the beginning it tends to be too conservative.

![Figure 4.8 Survival Plots of KD, KM and True](image)

4.3.2 Statistical model validation for noncensored cancer data parametric data validation analysis

It is well known that parametric analysis will result in the best estimate of the probability survival curves. However, if we can not justify parametric analysis, then we must proceed with nonparametric estimates using KD, KM or semi-parametric Cox PH survival models.
For the breast cancer data, we have survival times of 641 breast cancer patients with 48 being uncensored. Proceeding with goodness of fit statistical methods, using the Kolmogorov-Smirnov test we have identified that the two parameter gamma probability distribution fits the breast cancer data quite well. The gamma probability failure distribution with the maximum likelihood estimates of the shape and scale parameter is given by equation 4.18, below,

\[
\hat{f}(t, \hat{\alpha}, \hat{\beta}) = \begin{cases} 
  t^{0.997} e^{t/1.603} & \text{if } \frac{t}{1.603} \geq 3.997 (1.997), \\
  0 & \text{otherwise}
\end{cases} \quad (4.18)
\]

A graphical form of the gamma survival model is given below by Figure 4.9.

![Gamma Fitted Probability Density Function for The Cancer Data](image)

**Figure 4.9 Gamma Fitted Probability Density Function for The Cancer Data**

Thus, we can proceed to obtain the parametric survival function and proceed to develop and use it as reference to compare it with the KD, KM and Cox PH survival models.

The estimated parametric survival function based on the gamma pdf is given by

58
\[
\hat{S}_p(t, \hat{\cdot}, \hat{\cdot}) = 1 - t^{(1.997, \frac{t}{1.603})},
\]
and its corresponding estimate of the cumulative hazard function is given by
\[
\hat{H}_p(t, \hat{\cdot}) = \log(1 - t^{(1.997, \frac{t}{1.997})}).
\]

A graphical presentation of \( \hat{S}_p(t) \) is shown below by Figure 4.10.

![Graph of Parametric Survival Model](image)

**Figure 4.10 Plot of Parametric Survival Model**

**Kaplan-Meier Survival Analysis**

The estimate of the survival function of the Kaplan-Meier survival model is given by
\[
\hat{S}_{KM}(t) = \prod_{t_i < t} \frac{ni - di}{ni},
\]
and the corresponding estimate of the cumulative hazard function is given by
\[
\hat{H}_{KM}(t) = -\ln\left(\prod_{t_i < t} \frac{ni - di}{ni}\right),
\]
where \( n_i \) is the number of patients at risk just prior to time \( t_i \) and \( d_i \) is the number of deaths at exactly time \( t_i \).

A graphical display of the estimated KM survival curve and estimated KM cumulative hazard curve along with 95% confidence limits is shown by Figures 4.11 and 4.12.

From Figures 4.11 and 4.12 we can observe that the KM method is close to the parametric plot; however, since it is a step like function, one will prefer the parametric estimate which is a smooth curve.

![Survival Curves for Parametric and KM Models with 95% C.I.](image)

Figure 4.11 Survival Curves for Parametric and KM Models with 95% C.I.
Figure 4.12 The Cumulative Hazard Plot for Parametric and KM Method

Since the probability residual analysis follows the same procedure discussed in the previous section, we can conclude that the mean of the probability residuals is 0.00949, sample variance is .0196 with sample standard deviation is 0.14014 and standard error is 0.02023. Thus, the KM method is quite close to the parametric method.

The estimated value of the KD survival function is given by

$$\hat{S}_{KD}(x) = 1 - \frac{1}{x} \left( \frac{N}{\text{min}(SD(x), IQR(x))} \right)^{1.34n^{\frac{1}{3}}} \left( 1 - \frac{x}{1.06 \text{min}(SD(x), IQR(x))} \right)^{\frac{3}{4}} \left( \frac{x}{1.06 \text{min}(SD(x), IQR(x))} \right)^{\frac{3}{4}}, \quad (4.23)$$

and the estimate of the hazard function is given by
\[
\hat{h}_{KD}(x) = \frac{1}{N \cdot 1.06 \min(SD(x), IQR(x))} \left( \frac{3}{4} \left( 1 - \frac{x - x_i}{1.06 \min(SD(x), IQR(x))} \right)^2 \right) \cdot \frac{1}{1.34 n^{\frac{1}{3}}}.
\]

A graphical presentation of the estimated KD survival curve and estimated KD cumulative hazard curve is given by Figures 4.13 and 4.14, below, along with the parametric results.

From the Figures 4.13 and 4.14 we can observe that the KD curves for survival and cumulative hazard are closer to the parametric curves, which indicates that the KD survival method seems better than the KM survival method.

**Probability Residual Analysis**

The mean of the probability residual is 0.005529787, sample variance is 0.000553 with sample standard deviation is 0.02352216 and standard error is 0.003395132. Clearly, the KD method gives better estimates than the KM method in terms of both sample mean and sample standard error.

![Figure 4.13 The Parametric and KD Survival Curve](image)
Figure 4.14 The Cumulative Hazard Plot for Parametric and KD Method

Cox PH Model

The Cox PH model is a very popular semi-parametric method that has been used extensively to estimate the survival probability function of a given set of data that characterizes the failure time of a given patient.

Analytical Form

Hazard function of the Cox PH model is given by

\[ h_i(x) = h_0 \exp( \sum x_{it} + \sum x_{it} + \ldots + \sum x_{it} ), \]

and its survival function is given by

\[ S_i(x) = \exp( - \int h_0 \exp( \sum x_{it} + \sum x_{it} + \ldots + \sum x_{it} ) dt ), \]

where \( h_0 \) is the baseline hazard and all betas are the coefficients of the covariates.

The data set contains these covariates as shown in Table 4.4, below.
Table 4.5 The data set variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>stnum</td>
<td>Patient ID</td>
</tr>
<tr>
<td>tx</td>
<td>Randomized treatment: T=tamoxifen, B=radiation + tamoxifen</td>
</tr>
<tr>
<td>pathsize</td>
<td>Size of the tumour in cm</td>
</tr>
<tr>
<td>hist</td>
<td>Histology: DUC=Ductal, LOB= Lobular, MED= Medullary, MIX= Mixed, OTH= Other</td>
</tr>
<tr>
<td>hrlevel</td>
<td>Hormone receptor level: NEG=Negative, POS=Positive</td>
</tr>
<tr>
<td>hgb</td>
<td>Haemoglobin in g/l</td>
</tr>
<tr>
<td>nodediss</td>
<td>Whether axillary node dissection was done: Y=Yes, N=No</td>
</tr>
<tr>
<td>age</td>
<td>Age in years</td>
</tr>
</tbody>
</table>

From Table 4.5, we start with covariate of the first order, namely, stnum, tx, pathsize, hist, hrlevel, hgb, nodediss, age and all possible 2nd term and interactions. By using a stepwise selection with minimum Akaike's information criterion (AIC), we obtain the final model with only two first order terms that are significant, namely, tx and stnum. The Cox model summary is given by Table 4.5, below.

Table 4.6 Cox model summary

|       | coef     | exp(coef) | se(coef) | z      | Pr(>|z|) |
|-------|----------|-----------|----------|--------|---------|
| txT   | 1.660047 | 0.190130  | 0.763126 | 2.175  | 0.0296  |
| stnum | 0.005599 | 1.005615  | 0.002348 | 2.384  | 0.0171  |

Furthermore, the quality of the selected model based on the three statistical criteria is given in the Table 4.6, below, that support the quality of the fitted model.
Table 4.7 Cox model’s significance

<table>
<thead>
<tr>
<th>test</th>
<th>value</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood ratio test</td>
<td>5.39</td>
<td>2</td>
<td>0.06746</td>
</tr>
<tr>
<td>Wald test</td>
<td>5.69</td>
<td>2</td>
<td>0.05822</td>
</tr>
<tr>
<td>Score (logrank) test</td>
<td>5.74</td>
<td>2</td>
<td>0.05684</td>
</tr>
</tbody>
</table>

The fitted survival function is given by,

\[ \hat{S}_{\text{Coxph}}(x) = \exp \left( h_o \exp \left( 1.66x_{11} + 0.005599x_{12} \right) dt \right) \]  \hspace{1cm} (4.27)

with the corresponding estimate of the hazard function

\[ \hat{H}_{\text{Coxph}}(x) = h_o \exp \left( 1.66x_{11} + 0.005599x_{12} \right) dt. \] \hspace{1cm} (4.28)

From Tables 4.6 and 4.7 we can conclude that Cox PH model does not seem to fit the data very well because the p-value of the Likelihood ratio test (LRT), Wald test and Score test are all greater than .05.

From Figures 4.15 and 4.16 we can see that the Cox PH curve does not seem to fit the data very well.

Table 4.8 Residual analysis for uncensored data

<table>
<thead>
<tr>
<th>Methods</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>Rank of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox PH vs. fitted parametric</td>
<td>0.0095</td>
<td>0.1401</td>
<td>0.02023</td>
<td>3</td>
</tr>
<tr>
<td>Kaplan-Meier vs. fitted parametric</td>
<td>0.0095</td>
<td>0.1401</td>
<td>0.0202</td>
<td>2</td>
</tr>
<tr>
<td>Kernel density vs. fitted parametric</td>
<td>0.00553</td>
<td>0.0235</td>
<td>0.003395</td>
<td>1</td>
</tr>
</tbody>
</table>

From Table 4.8 we can conclude that if we exhaust all possible parametric choices, then we have to perform nonparametric analysis. In this situation we will recommend the KD method in terms of its smoothness and less standard error from the probability residuals analysis.
Figure 4.15 The Survival Plot for Parametric and Cox PH Model with 95% CI

Figure 4.16 The Cumulative Hazard Plot for Parametric and Cox PH Model
A graphical comparison of survival plots for KD, parametric, KM and Cox PH model are summarized by Figure 3.9, below,

![Graph showing survival curves for KD, Parametric, KM, and Cox PH models]

Figure 4.17 The Survival Curves for KD, Parametric, KM and Cox PH Model

From a visual evaluation of the four survival curves, we can conclude that the KD model comes the closest to the parametric model followed by KM and Cox PH. Thus, using the real cancer data we can conclude that the KD is better than the KM and Cox PH models. This is consistent with MC simulation that we initially performed. It is recommended that when the censoring rate is small then KD is the best model to use.
4.3.3 Statistical model validation for censored data

Quite often we deal with censored data due to limited and difficult experimental conditions. In the present study we are interested in investigating how KD analysis performs under a censored data situation. The problem is that we will never know the true state of nature under the censored circumstances and the only information we are certain of is that by the time it is censored the patient is still alive.

How to conduct a goodness of fit test for censored data is still an open problem. Edel A. Pena, [9], discussing the subject matter stated that we can only reject some probability distributions but still can not have the unique best distribution to probabilistically characterize the censored data. In this study we will perform KM, KD, Cox PH and parametric survival analysis for censored data and evaluate their response.

Parametric Survival Analysis

Despite the difficulties of the censored data, we find the best possible fit for the cancer data follows two parameter Weibull distribution.

The two parameter Weibull pdf is given by 4.1, below,

\[ f(x; \alpha, \beta) = -\frac{x^\alpha}{\beta^\alpha} \exp \left( -\frac{x}{\beta} \right), \quad \beta > 0, \quad x \geq 0 \quad \text{and} \quad x < 0 \quad \text{otherwise} \quad (4.29) \]

where \( \alpha \) and \( \beta \) are the shape and scale parameter respectively.

The cumulative distribution function (CDF) of 4.1 is given by

\[ F(x; \alpha, \beta) = 1 - \exp \left( -\frac{x}{\beta} \right), \quad \beta > 0, \quad x \geq 0 \quad (4.30) \]

The survival function of the Weibull distribution is given by

\[ S(x; \alpha, \beta) = \exp \left( -\frac{x}{\beta} \right), \quad \beta > 0, \quad x \geq 0 \quad (4.31) \]
and the hazard function is of the form

$$h(x) = - \left( \frac{x}{\lambda} \right)^{\lambda - 1}, \quad > 0, \quad > 0, \quad x \geq 0.$$  \hspace{1cm} (4.32)

From the parametric analysis, we obtain the estimates \( \hat{\lambda} = 1.0962 \) and \( \hat{\mu} = 59.243 \). The estimated parametric survival function is given by

$$\hat{S}(t, \hat{\lambda}, \hat{\mu}) = \exp \left( - \frac{x}{\hat{\mu}} \right)^{\hat{\lambda}} \cdot \frac{x}{\hat{\mu}}, \quad x \geq 0 \hspace{1cm} (4.33)$$

We should assume that this parametric survival model will be the best possible survival model for this censored cancer data. We will utilize this survival model to evaluate the performance of the other three nonparametric survival methods, namely, KM, Cox PH and KD survival methods.

**KM Survival Analysis**

The survival probability estimate of the censored data using the KM model with the parametric survival model is given by Figure 4.18, below. In comparing the KM survival curve with the parametric model using the probability residuals, we have found that the mean of the probability residual is 0.000495, sample variance is 4.596\times10^{-5}, sample standard deviation is 0.0068 and standard error is 0.0024.

![Figure 4.18 The Survival Plot for KM Model](image-url)
Cox PH Model

To select the best possible Cox PH model for censored data, we consider the model has all terms significant with the minimum AIC. Through statistical testing we have found that six first order terms and two interactions significantly contribute to the response variable. These attributable variables are tx, pathsize, nodediss, age, hrlevel, stnum, tx:age and nodediss:hrlevel. Thus, for the subject data and the attributable variables using the Cox PH model we plot the probability survival curve with the parametric survival curve and they are shown by Figure 4.19, below. In comparing the Cox PH curve with the estimated parametric survival curve, we found the mean of the probability residual is 0.028, sample variance is 0.000347, sample standard deviation is 0.0186 and standard error is 0.0066.

![Figure 4.19 The Survival Plot for Cox PH Model](image)

KD survival analysis

Despite the difficulties in working with censored data, we proceeded to use the nonparametric KD procedure to estimate the survival curve together with the fitted parametric survival curve. The results are shown by Figure 4.20, which are
different than what we have found using KM and Cox PH models in terms of the smoothness of the curve.

In comparing the KD survival curve with the fitted parametric model, we have found the mean of the probability residual is 0.00883, sample variance is 2.942*10^{-5}, sample standard deviation is 0.0054 and standard error is 0.00191.

Table 4.9, below, summarizes the response of the three survival analysis models, KM, Cox PH and KD, in comparison with the parametric model using the two parameter Weibull probability density function to characterize the failures. Thus, if we assume that we can proceed to statistically analyze the censored data, all three survival models performed well, but the edge goes to the KD model in terms of the smaller sample variance and standard error.

![Figure 4.20 The Survival Plot for KD Model](image)

71
Table 4.9 Residual analysis for censored data

<table>
<thead>
<tr>
<th>Methods</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>Rank of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox PH vs fitted parametric</td>
<td>0.028</td>
<td>0.0186</td>
<td>0.00658</td>
<td>3</td>
</tr>
<tr>
<td>Kaplan-Meier vs fitted parametric</td>
<td>0.0005</td>
<td>0.0068</td>
<td>0.0024</td>
<td>2</td>
</tr>
<tr>
<td>Kernel density vs fitted parametric</td>
<td>0.00883</td>
<td>0.0054</td>
<td>0.00191</td>
<td>1</td>
</tr>
</tbody>
</table>

4.4 Conclusions

The present study consists of three parts in comparing the effectiveness of three survival analysis models, namely, KM, Cox PH and KD.

Initially, using Monte Carlo simulation we compare the subject models with parametric survival models and found that the proposed KD survival model gives as good results, if not better, than the KM.

The second part consists of using actual uncensored breast cancer data. Performing a similar evaluation, the results support that the proposed KD model gives results in better estimates than the popular KM and Cox PH models with interactions.

Thirdly, we performed the same analysis with actual censored breast cancer data. Although working with censored data is quite difficult to justify such an analysis, under the circumstances we analyzed the data and the results are similar to the Monte Carlo simulation and using the uncensored data.
Chapter 5

Identify Attributable Variables and Interactions in Breast Cancer

5.1 Background and Data

National cancer institute (2010) defines breast cancer as following: Cancer that forms in tissues of the breast, usually the ducts (tubes that carry milk to the nipple) and lobules (glands that make milk). It occurs in both men and women, although male breast cancer is rare.

The proposed model that we are developing includes individual variables, interactions, and higher order variables if applicable. In developing the statistical model, the response variable is the tumor size at diagnosis for breast cancer patients. We have identified 25 possible attributable variables for breast cancer, denoted, X1, X2,..., X25. For example, X1 stands for patient’s age at diagnosis and X2 stands for the patient’s Year of Birth. The list of attributable variables is in Table 5.1, below. In this study, we would like to find the relation between the tumor size and all other attributable variables. We cannot use survival time to predict the tumor size since death time happens after the tumor is detected. Therefore, we exclude the variable survival time(x24) and the censoring indicator function vss (x25) in the first part of study. Thus, we have only 23 variables left to construct our statistical model from Xu and Tsokos (2011).

In the present analysis, we used real data from the Surveillance Epidemiology and End Results (SEER) Program. SEER collects and compiles information on
incidence, survival, and prevalence from specific geographic areas representing about 26 percent of the U.S. population plus cancer mortality for the entire U.S.

The proposed statistical model is useful in predicting the tumor size given data for the attributable variables. It is statistically evaluated using R square, R square adjusted, the PRESS statistic and several types of residual analyses. Finally, its usefulness is illustrated by utilizing different combinations of the attributable variables.

In addition, the attributable variables are ranked according to their contributions to accurately estimate a patient’s tumor size.

5.2 Methodology

We randomly extract 155 uncensored breast cancer patients’ information from SEER data base. The data was obtained from 2000 to 2006. We want to develop a statistical model with full information instead of censored; therefore, we will use the 155 uncensored patients’ information to construct our statistical model. The data tree diagram is shown by Figure 5.1, below.

We proceed to develop a statistical model taking into consideration the twenty four attributable variables listed in Table 5.1. The form of the statistical model is given by tumor size as a function of \( (x_1, x_2, \ldots, x_{23}) \). Note that some of the variables’ values are obtained after the tumor size is recorded. In our analysis all the patients in the data base have breast cancer. We utilize the values of the tumor size once the patient has gone through a diagnostic process. Thus, the general statistical form of the proposed model with all possible attributable variables and interactions will be of the form in equation 5.1.

\[
TS = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \ldots + \beta_i A_i + \beta_1 B_1 + \beta_2 B_2 + \ldots + \beta_j B_j 
\]  

(5.1)
Table 5.1 List of attributable variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Full name of variables</th>
<th>Short form</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Age at diagnosis</td>
<td>age</td>
</tr>
<tr>
<td>X2</td>
<td>Year of Birth</td>
<td>birthy</td>
</tr>
<tr>
<td>X3</td>
<td>Birth Place</td>
<td>birthp</td>
</tr>
<tr>
<td>X4</td>
<td>Sequence Number Central</td>
<td>snc</td>
</tr>
<tr>
<td>X5</td>
<td>Month of diagnosis</td>
<td>month</td>
</tr>
<tr>
<td>X6</td>
<td>Year of diagnosis</td>
<td>year</td>
</tr>
<tr>
<td>X7</td>
<td>Primary Site</td>
<td>ps</td>
</tr>
<tr>
<td>X8</td>
<td>Laterality</td>
<td>la</td>
</tr>
<tr>
<td>X9</td>
<td>Histologic Type ICDO3</td>
<td>ht</td>
</tr>
<tr>
<td>X10</td>
<td>Behavior Code ICDO3</td>
<td>bc</td>
</tr>
<tr>
<td>X11</td>
<td>Type of Reporting Source</td>
<td>trs</td>
</tr>
<tr>
<td>X12</td>
<td>RXSumm SurgPrimSite</td>
<td>rxps</td>
</tr>
<tr>
<td>X13</td>
<td>RXSumm Radiation</td>
<td>rxr</td>
</tr>
<tr>
<td>X14</td>
<td>RXSumm RadtoCNS</td>
<td>rxcns</td>
</tr>
<tr>
<td>X15</td>
<td>Age Recode Year olds</td>
<td>ager</td>
</tr>
<tr>
<td>X16</td>
<td>Site Recode</td>
<td>sr</td>
</tr>
<tr>
<td>X17</td>
<td>CSS chema</td>
<td>css</td>
</tr>
<tr>
<td>X18</td>
<td>AJCC stage3 rdition</td>
<td>ajcc</td>
</tr>
<tr>
<td>X19</td>
<td>First malignant primary indicator</td>
<td>findi</td>
</tr>
<tr>
<td>X20</td>
<td>State-county recode</td>
<td>scr</td>
</tr>
<tr>
<td>X21</td>
<td>Race</td>
<td>race</td>
</tr>
<tr>
<td>X22</td>
<td>Cause of Death to SEER site recode</td>
<td>cod</td>
</tr>
<tr>
<td>X23</td>
<td>Sex</td>
<td>sex</td>
</tr>
<tr>
<td>X24</td>
<td>Survival time recode</td>
<td>survtime</td>
</tr>
<tr>
<td>X25</td>
<td>Vital Status recode</td>
<td>vss</td>
</tr>
</tbody>
</table>
Figure 5.1 Breast Cancer Data Tree Diagram

Here, TS stands for tumor size, the   and   are the coefficients, and A is the first order term of the attributable variables and B are the possible interactions and higher order terms. The object is to develop the most representative estimate of the above model based on available data.

One of the basic underlying assumptions in formulating an estimate of the above statistical model is that the response variable should be Gaussian distributed. Unfortunately, in the present form that is not the case. This fact is clearly demonstrated by the QQ plot shown by Figure 5.2, below.
Furthermore, the Shapiro-Wilk normality test (Shapiro and Wilk 1965) with the necessary calculation of the test statistic $W = 0.7437$ and $p$-value = $3.787 \times 10^{-15}$ is additional evidence that the tumor size does not follow normal probability distribution. We proceed in utilizing the Box-Cox transformation (Box and Cox 1964) to the tumor size to determine if such a filter will modify the given data to follow the normal distribution so that we can proceed to formulate the proposed statistical model. Applying the Box-Cox transformation results in the statistical information presented in Table 5.2. One tumor size data’s value is zero and Box-Cox transformation can only apply to a positive data set. Therefore, we use $0.000000000000001$ to replace this zero value so we can perform Box-Cox transformation.

Therefore, we decide to use the transformed tumor size as our response and we redo the Box-Cox transformation test with the QQ plot to see if the transformed data will follow Gaussian probability distribution. With the Box-Cox results for
transformed data presented in Table 2, we can conclude that the transformed data follows the Gaussian probability distribution. Also the QQ plot in Figure 5.3 supports the fact that the transformed tumor size follows the Gaussian probability distribution. Since the transformed power is only .2659, we decide the logarithmic filter will be appropriate in this situation for the transformed tumor size data.

Table 5.2 Box-Cox Transformation for Normality for the Original Transformed Data

<table>
<thead>
<tr>
<th></th>
<th>Est.Power</th>
<th>Std.Err.</th>
<th>Wald(Power=0)</th>
<th>Wald(Power=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Data</td>
<td>0.2659</td>
<td>0.0339</td>
<td>7.8445</td>
<td>-21.6563</td>
</tr>
<tr>
<td>Transformed Data</td>
<td>1</td>
<td>0.1275</td>
<td>7.8444</td>
<td>-2e-04</td>
</tr>
</tbody>
</table>

Thus, we can proceed to estimate the coefficients of the attributable variables for the filtered transformed tumor size data to obtain the coefficient of all possible interactions and at the same time determine the significant contributions of both attributable variables and interactions.

![Normal Q-Q Plot](image)

*Figure 5.3 QQ Plot for Testing Normality for Transformed Data*
We begin with the previously defined twenty-four attributable variables \( x_1, x_2, \ldots, x_{24} \) and the two hundred and seventy-six first degree 2nd order interaction between each pair, and the two thousand and twenty-four first degree 3rd order interactions between any three variables. We did not consider any 4th and higher order interactions. Since we already have a lot more terms than patients in the data itself, we utilized sub modeling skills and two way selection procedures to construct our model. To develop the models, initially we start building our model with a total of two thousand and three hundred and twenty-four terms that include initial contribution of the attributable variables and the described interactions. More than thirty candidate models were constructed.

During our statistical analysis in the estimation process, we found only three of the twenty-four attributable variables were significant contributors. We found only three higher order interactions to significantly contribute. The significantly contributing and interaction variables are RXR(X14), RXPS(X13) and AJCC(X19). However, SNC(X5), HT(X10), themselves individually do not significantly contribute to the response variables but when they interact with other variables they do significantly contribute to the response variable. Therefore, we still keep them in our final model. There are thirty-one missing values in the variable AJCC. We use the mean of the rest of the data value in the variable AJCC to replace the NA value in order to perform prediction of the model. Thus, the results of estimation of equation 1 are given by equation 2 as follows

\[
\ln(\hat{T}S^{2659}) = 2.7 - 2.09 \times 10^{-2} \times X_5 + 2.14 \times 10^{-4} \times X_{10} + 3.28 \times 10^{-3} \times X_{13} - 3.3 \times X_{14} + 2.72 \times 10^{-3} \times X_{19} + .24 \times X_{14} \times X_{19} + 1.91 \times 10^{-4} \times X_5 \times X_{10} \times X_{14} + 1.19 \times 10^{-4} \times X_5 \times X_{14} \times X_{19}
\]

(5.2)

We will utilize the initial transformation that we used to transform the response data to get the result in equation 3 by taking the exponential on both side of the 2 and then take 3.61’s power on both sides of the equation 2.
\[ \hat{TS} = (\exp(2.7 - 2.09 \ 10^{-2}X_4 \ 2.14 \ 10^{-4}X_{10} + 3.28 \ 10^{-3}X_{13} - 3.3X_{14} \\
+ 2.72 \ 10^{-3}X_{19} + .24X_{14} \ X_{19} + 1.91 \ 10^{-2}X_5 \ X_{19} \ X_{14} \\
\times 1.19 \ 10^{-1}X_3 \ X_{14} \ X_{19}))^{1.76} \] (5.3)

The proposed statistical model's high quality has been evidenced by R square and R square adjusted, which are the key criteria of evaluating such models. The regression sum of squares (SSR), also called the explained sum of squares, is the variation that is explained by the regression model. The sum of squared errors (SSE), also called the residual sum of squares, is the variation that is left unexplained. The total sum of squares (SST) is proportional to the sample variance and equals the sum of SSR and SSE. The coefficient of determination, \( R^2 \), is defined as the proportion of the total sum of squares that is explained by the model. That is, \( R^2 = \) SSR/SST. It provides an overall measure of how well the model fits the data. R-square adjusted will adjust for the degrees of freedom in the model and it works better when there are a substantial number of parameters in the model. R-square adjusted is given by:

\[ R^2_{adj} = 1 - \frac{SSE/df(SSE)}{SST/df(SST)} \]

Here df(SSE) means the degree of freedom (DOF) of the SSE and df(SST) means the DOF of SST.

The prediction of residual error sum of squares (PRESS) statistics evaluate how good the estimation is each time a data point is removed. The PRESS statistic is defined by:

\[ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y - \hat{y}_i}{H_{ii}} \right)^2 \]

(Allen 1971 and 1974) and Hii is the diagonal elements of the projection matrix. Therefore, we should choose the model with the smallest PRESS statistics from all candidate models.

For our final model the R squared is 0.88 and R squared adjusted is 0.87. Both R squared value and R squared adjusted value are high (close to 90%) and these two are very close to each other. This shows our model's R squared increase is not due to the increase of the parameters' estimates, but rather the good quality
the proposed model to predict tumor size given values of the identified attributable variables. Secondly, the PRESS statistics’ results support the fact that the proposed model is of high quality. We list in Table 5.3 the best three models based on the PRESS statistic out of total thirty-six models. From that table it is clear that the best model is number 36, which is our final model.

Furthermore, R square and R square adjusted are calculated for those 36 models which are of interest but the proposed model still gives the best possible estimates of the tumor size for breast cancer in SEER’s data.

<table>
<thead>
<tr>
<th>Model number</th>
<th>PRESS value</th>
<th>Rank of the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>96.73797</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>98.4167</td>
<td>2</td>
</tr>
<tr>
<td>36</td>
<td>104.8218</td>
<td>1</td>
</tr>
</tbody>
</table>

In Table 5.4 we have listed all the important attributable variables and interactions.

For example, X5 and X10 are not significant by themselves, only in combination with the others. We summarize the attributable variables individually and interactively in the following schematic network as showed in Figure 5.4.

In Table 5.5 are the ranks of the most important attributable variables with respect to their contribution to estimating tumor size. The interaction among X5, X14 and X19 ranks number one. This is the interaction among the sequence number central(SNC), RXsumm Radiation(RXR) and AJCC stage3 rdedition(AJCC).
### Table 5.4 List of Attributable Variables

<table>
<thead>
<tr>
<th>No</th>
<th>Individual variables</th>
<th>Name of individual variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X5</td>
<td>Sequence Number Central</td>
</tr>
<tr>
<td>2</td>
<td>X10</td>
<td>Histologic Type ICDO3</td>
</tr>
<tr>
<td>3</td>
<td>X13</td>
<td>RXSumm SurgPrimSite</td>
</tr>
<tr>
<td>4</td>
<td>X14</td>
<td>RXSumm Radiation</td>
</tr>
<tr>
<td>5</td>
<td>X19</td>
<td>AJCC stage3 rедакция</td>
</tr>
<tr>
<td>6</td>
<td>X14:X19</td>
<td>RXR ∩ AJCC</td>
</tr>
<tr>
<td>7</td>
<td>X5:X10:X14</td>
<td>SNC ∩ HT ∩ RXR</td>
</tr>
<tr>
<td>8</td>
<td>X5:X14:X19</td>
<td>SNC ∩ RXR ∩ AJCC</td>
</tr>
</tbody>
</table>

### BREAST CANCER: TUMOR SIZE

![Breast Cancer Attributable Variable Diagram](image)

**Figure 5.4 Breast Cancer Attributable Variable Diagram**
Table 5.5 Rank of Variable According to Contributions

<table>
<thead>
<tr>
<th>Rank</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X5:X14:X19</td>
</tr>
<tr>
<td>2</td>
<td>X14:X19</td>
</tr>
<tr>
<td>3</td>
<td>X19</td>
</tr>
<tr>
<td>4</td>
<td>X5:X10:X14</td>
</tr>
<tr>
<td>5</td>
<td>X5</td>
</tr>
<tr>
<td>6</td>
<td>X14</td>
</tr>
<tr>
<td>7</td>
<td>X13</td>
</tr>
<tr>
<td>8</td>
<td>X10</td>
</tr>
</tbody>
</table>

5.3 Validation of The Proposed Model

We use two methods to validate the model. The first method is to use the proposed model to calculate the predicted value for each tumor size and then calculate the residuals. A residual is defined as the original value minus the predicted value. Table 5.6 shows the last ten residuals out of the total one hundred fifty-five residuals.

The mean of the residuals is -.0286, variance of the residuals is 1.588, standard deviation is 1.26 and standard error of the residuals is .1012.

The second method we will utilize is called cross validation. The basic idea is we will save some portion of the data for validation. We construct our model using only the data left and the constructed model will be same structure as our proposed model with only coefficients being different. We will test the quality of the model using three settings.

We first randomly divide the data into two datasets of the same size. We use one of the datasets to construct the model and then use the resulting model to predict the values in the other dataset. Then we will switch the two data sets and repeat the procedure. The mean of all residuals turned out to be 1.0652916.

Next, we divided the dataset into six small data sets and use five of them to construct the model and validate the model using the sixth one. We will repeat the
same procedure for each of the six small datasets. The mean of all residuals was 0.1318486.

Finally, we divided the dataset into 155 datasets and use all 154 datasets to construct the model and validate the model using the one left out. We repeat the procedure 155 times. Table 5.6 shows the last ten residuals out of the total one hundred fifty-five residuals.

The mean of the residuals was 0.634, the variance of the residuals was 42.89, standard deviation of the residuals was 6.55 and standard error of the residuals was 0.53.

<table>
<thead>
<tr>
<th>No</th>
<th>Original data Residual Values</th>
<th>Cross Validation Residual Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>146</td>
<td>1.3264266</td>
<td>0.13328523</td>
</tr>
<tr>
<td>147</td>
<td>1.0579828</td>
<td>0.08479963</td>
</tr>
<tr>
<td>148</td>
<td>0.9756659</td>
<td>0.07219656</td>
</tr>
<tr>
<td>149</td>
<td>1.7362950</td>
<td>0.36801567</td>
</tr>
<tr>
<td>150</td>
<td>0.9643773</td>
<td>0.07578824</td>
</tr>
<tr>
<td>151</td>
<td>1.2427113</td>
<td>0.11712640</td>
</tr>
<tr>
<td>152</td>
<td>1.3705402</td>
<td>0.14230504</td>
</tr>
<tr>
<td>153</td>
<td>1.3640997</td>
<td>0.14117093</td>
</tr>
<tr>
<td>154</td>
<td>1.6072370</td>
<td>0.19570259</td>
</tr>
<tr>
<td>155</td>
<td>1.9573079</td>
<td>0.29023850</td>
</tr>
</tbody>
</table>

5.4 Conclusions

We can conclude from our extensive statistical analysis that there are only three significant attributable variables to the tumor size for breast cancer namely, RXR(X14), RXPS(X13) and AJCC(X19). As for SNC(X5), HT(X10). They themselves individually do not significantly contribute to the response variables; however, when they interact with other variables, they do significantly contribute to the response variable. Furthermore, we also tested two thousand and three hundred possible
interactions of the attributable variables and we found three interactions to significantly contribute to tumor size for breast cancer.

This model is useful for a number of reasons.

1. One can also use the proposed model to generate various scenarios of the breast cancer tumor size as a function of different values of the subjective entities for data simulation purpose.

2. It can be used to identify the significant attributable variables.

3. It identifies the significant interactions of these attributable variables.

4. The significant contributions to the breast cancer tumor size are ranked.

5. A confidence interval for the tumor size can be constructed with parametric analysis. By obtaining the $100\times (1-\alpha)$% confidence limits for the response, we can describe how confident we are that our estimate is close to the actual tumor size.

6. The model as shown in equation 3 can be used to perform surface response analysis to place the restrictions on the significant attributable variables and interactions to minimize the breast cancer tumor size. We can also put restrictions on the variables to minimize the response of the tumor size by nonlinear control with $100\times (1-\alpha)$% confidence limits.

In the present study, we performed parametric analysis to estimate tumor size for breast cancer patients for data simulation purpose. The initial measurement of tumor size was collected from the SEER database. Those data do not follow normal probability distribution. Using the standard Box-Cox transformation, the SEER tumor size data became approximately normally distributed. We developed a “nonlinear” statistical model (nonlinear in terms of the power and logarithm of the response variable). Through the process of developing the statistical model, we found only four variables, namely, rrx(X14), rxs(X13) and ajcc(X19) and three interactions that significantly contribute to the tumor size. The proposed statistical model was evaluated using the R-square, R-square adjusted, PRESS statistics and three cros

85
validation methods, all of which support the high quality of the developed statistical model. This model can be used to obtain a good estimate of tumor size knowing the four significantly attributable variables and three interaction terms.

We validate this model on four different data sets. Since this type of model is strongly data driven, we need to make the necessary modification. For example, the Box-Cox transformation's power estimation needs to be changed according to the data set. The specific model's interaction and estimation of the coefficients need to be recalculated. After the necessary modifications these four data sets fit the model as following. The model fits one of the data sets pretty well with high R-square more than 90. The model fit two data sets not very well with R-square around 30 and one data set pretty bad with R-square around 10. Considering the complexity and random behavior of the breast cancer and the limitation of our available data set, we are not surprised with the result. We wish to make the update of the model based on large scale data set for the future study.
Chapter 6

Power Law Process for Evaluating Stage I & II Ductal Breast Cancer Treatment

6.1 Background and Data

In the present study we utilize the Power Law process (PLP), also known as Non-homogenous Poisson Process and Weibull Process, to evaluate the effectiveness of a given treatment for Stage I & II Ductal breast cancer patients. We study the behavior of the shape parameter of the intensity function to evaluate the response of a given treatment with respect to its effectiveness for the subject cancer.

Breast cancer (malignant breast neoplasm) is cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk, from Sariego (2010).

In the present analysis, we use real data from Surveillance Epidemiology and End Results (SEER) Program to test the proposed model. U.S. National Institutes of Health (NIH) (2010) collects information on incidence, survival, and prevalence from specific geographic areas representing 26 percent of U.S. population and compiles reports on several type of cancer and includes mortality rate for the entire U.S. in SEER program.


The schematic diagram, Figure 6.1, gives a complete picture of the data base that we are using in the present study. We randomize our data set to reduce the random errors by performing simple random sampling procedures. We randomly extract 500,000 breast cancer patients’ information from a total of 578,134 SEER data base. Out of the 500,000 breast cancer patients we have 496,783 female patients and 3,217 male patients. The female patients are broken into different race groups, white, African-American and Asian (includes others). Thus, we have 426,302 white patients, 39,681 African-American patients, 29,015 Asian patients and 1,785 unspecified patients. Within each patient’s group we have four types of breast cancer: Ductal, Medullary, Lobular and other (unspecified). For each type of breast cancer, we further divide patients according to their American Joint Committee on Cancer (AJCC) Cancer Staging such as stage I, II, III, IV and others. It is a commonly occurring cancer in white females than other races. In the present study we concentrate in using only the cancer data for white females. Ductal is a major type of breast cancer for white females. Thus, we will focus on ductal breast cancer for the white race in this study.
The breast cancer data tree

Original SEER data of 578,134 patients with breast cancer

500,000 patient data which randomly choose from SEER data

496,783 female patients

3,217 male patients

1,785 unknown (unspecified) patients

426,302 white patients

39,681 African American patients

29,015 Asian (other) patients

295,362 Ductal patients

63,968 Lobular patients

27,845 Ductal patients

4,767 Lobular patients

6,049 Other patients

61,695 Other patients

22,360Ductal patients

2,917 Lobular patients

3,396 Other patients

5,277 Medullary patients

1,020 Medullary patients

342 Medullary patients

67,450 Patients in WD Stage I

50,453 patients in WD Stage II

7,206 patients in WD Stage III

4,781 patients in WD Stage IV

Where WD stage I stands for White Ductal cancer patients in AJCC Stage I

Figure 6.1 Breast Cancer Data Tree Diagram
White Ductal Cancer Patients in Stage I

WD stage I stands for white ductal cancer patients in AJCC Stage I. Similarly, WD stage II, III and IV stand for white ductal cancer patients in AJCC Stage II, III and IV.

For white ductal patients in stage I, we divide them into two groups, patients who are still living and patients who are deceased as shown below in Figure 6.2. For those patients who are deceased, we break it down into patients who are deceased because of breast cancer and patients who are deceased because of other reasons. For those patients who are deceased because of breast cancer, we have different treatment information for those patients. We will construct PLP with respect to white ductal patients in stage I to compare the different treatments’ effects.

Figure 6.2 Breast Cancer Data Tree Diagram white Ductal Stage I patients
White Ductal cancer patients in Stage II

For white ductal patients in stage II, we divide them into two groups, patients who are still living and patients who are deceased, as shown below in Figure 6.3. For those patients who are deceased, we break it down into patients who are deceased because of breast cancer and patients who are deceased because of other reasons. For those patients who are deceased because of breast cancer, we have different treatments’ information for those patients. We will construct PLP with respect to white ductal patients in stage II to compare the different treatments’ effects.

Figure 6.3 Breast Cancer Data Tree Diagram white Ductal Stage II patients
The most commonly stage to clarify breast cancer patients are stages I and II. Thus, we shall consider these stages in the present study using the PLP to determine the effectiveness of the different treatments as shown in Figure 6.2 and 6.3, above.

6.2 Methodology

According to Tsokos (1997), Power law process also known as the non-homogeneous poisson process (NHPP) as well as Weibull process (WP). The PLP is also considered as a counting process. Let \( \{N(t), t \geq 0\} \) to be a counting process that possesses the following properties:

1. \( N(t) \geq 0 \)
2. \( N(t) \) is an integer.
3. If \( s < t \) then \( N(s) \leq N(t) \).

If \( s < t \), then \( N(t) - N(s) \) is the number of events occurred during the interval \((s, t]\).

A Poisson process is the stochastic process in which events occur continuously and independently of one another. The Poisson process is a collection \( \{N(t) : t \geq 0\} \) of random variables, where \( N(t) \) is the number of events that have occurred up to time \( t \) (starting from time \( 0 \)). The number of events between time \( a \) and time \( b \) is given as \( N(b) - N(a) \) and has a Poisson distribution. Each realization of the process \( \{N(t)\} \) is a non-negative integer-valued step function that is non-decreasing.

For PLP or NHPP, the rate parameter may change over time. In this case, the generalized rate function is given as \( \lambda(t) \). Now the expected number of events between time \( a \) and time \( b \) is

\[
\int_a^b \lambda(t) \, dt.
\]

(6.1)
Thus, the number of arrivals in the time interval \((a, b]\), given as \(N(b) - N(a)\), follows a Poisson distribution with associated parameter \(\lambda, a, b\)

\[
P[(N(b) - N(a)) = k] = \frac{e^{-\lambda} (\lambda b - a)^k}{k!} \quad k = 0, 1, \ldots
\] (6.2)

A homogeneous Poisson process may be viewed as a special case when \(\lambda(t) = \lambda, a\) constant rate.

NHPP has the intensity function:

\[
(\lambda) = \lambda_{0} e^{-\lambda_{0} t}, \quad \text{for} \quad t > 0, \quad \lambda > 0, t > 0.
\] (6.3)

The mean value function \(\mu(t)\) of the process is:

\[
\mu(t) = E(N(t)) = \int_{0}^{t} \lambda(s) ds = \int_{0}^{t} \lambda_{0} e^{-\lambda_{0} s} ds = \frac{t}{\lambda_{0}}
\] (6.4)

We know that if the parameter beta is greater than one in survival analysis, then the failure time increase means the survival rate decreased. If beta is less than one in survival analysis, then the failure time decrease which means the survival rate increase. If beta equals to one then the failure time is constant and the NHPP will become homogenous passion process (HPP) from Rigdon, S. E. and A. P. Basu (2010).

The unbiased estimator of beta is provided by Bain and Enelhardt (1991).

\[
\hat{\beta}_u = \frac{n}{n} \quad \hat{\beta}_{MLE} = \frac{n}{\log \left( \frac{t_{n}}{t_{1}} \right)}
\] (6.5)

is indicator function. If \(n = 1\) the system will be failure time truncated, means our system is restricted by a number of tails and we will stop the testing when we reach that number of testing. If \(n = 0\) the system will be time truncated, means our system is restricted by a final failure time and we will stop the testing when we reach that time.

93
\[ \gamma = \frac{t_n}{n} \]  

(6.6)

In our study, we belong to the first case that is we have fixed time of cases. We divide all patients into four groups according to their cancer stage. Within each stage, we have the information that the patients receive what kind of treatment or no radiation treatment at all. Therefore within each stage we divide the patients into four groups with respect the treatment they receive, namely without treatment, treatment 1, 2 and 3. Here treatment 1 refers to the beam radiation. Treatment 2 refers to radioactive implants treatment and treatment 3 means combination treatment. With the limit of our data sources we do not have too many patients choose treatment 2 and 3.

After we calculate the alpha and beta value for this PLP for each treatment then we will compare the results and observe the pattern. Since the white race is the major population here and Ductal patients is also the dominate type. We will focus on white doctal breast cancer patients in this study. The estimation of the parameter is shown in Table 6.1, below.

The information summarized as the graphical interpretation is given in Figure 6.4 below. We can observe the pattern for the key parameter beta. For example, \( \gamma_{11} \) is 1.11 which means if a patient does not receive any treatment, very likely, the patient’s condition will become worse because the indication of the growth of the tumor which will lead to the progression of cancer. It may lead to the patient from stage I to become stage II or even higher stages. If we look at \( \gamma_{31} \) and \( \gamma_{32} \), we can tell if a patient receives treatment 3 in stage I, then the patients will have a better result than getting the same treatment for the patients in stage II.
Figure 6.4 Evaluation Chain for PLP

We found that for those cases when beta are less than one it means the tumor size decreases which means our treatment for breast cancer works. From Table 6.1 we can conclude that for patients in early stage, say I and II, without treatment will surely increase the tumor size and speed the death time. For treatment 1: beam radiation works for stage I but not for stage II. For treatment 2: radioactive implants, it seems does not work well for both stage I and II and we do not have enough data to conduct the PLP for stage III and IV. For treatment 3: a combination of the treatments works well in stage I and II and we also lack of data to conduct PLP for stage III and IV.
Table 6.1 Parameter estimation for PLP

<table>
<thead>
<tr>
<th></th>
<th>Stage I</th>
<th>Stage II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha without treatment</td>
<td>94.5112</td>
<td>113.8267</td>
</tr>
<tr>
<td>Alpha with treatment 1</td>
<td>56.17724</td>
<td>92.982</td>
</tr>
<tr>
<td>Alpha with treatment 2</td>
<td>76.03755</td>
<td>66.60</td>
</tr>
<tr>
<td>Alpha with treatment 3</td>
<td>33.8427</td>
<td>41.35</td>
</tr>
<tr>
<td>Beta without treatment</td>
<td>1.110023</td>
<td>1.167756</td>
</tr>
<tr>
<td>Beta with treatment 1</td>
<td>0.9943948</td>
<td>1.1195</td>
</tr>
<tr>
<td>Beta with treatment 2</td>
<td>1.112772</td>
<td>1.076</td>
</tr>
<tr>
<td>Beta with treatment 3</td>
<td>0.8635</td>
<td>0.929</td>
</tr>
</tbody>
</table>

All the intensity function plots are in Figure 6.5 to Figure 6.12, below.

![Figure 6.5 Stage I Breast Cancer Intensity Function without Treatment](image)

Figure 6.5 Stage I Breast Cancer Intensity Function without Treatment

The graph given by Figure 6.5, above, shows that as the cumulative time of the patients increases, the intensity function will also increase, which indicates as
expected that the tumor size increases and cancer progresses. This result verifies the result we get from the estimate of the parameter $\beta_1$. The graph given by Figure 6.6, below, shows that as the cumulative time of the patients increases, the intensity function will also decrease which indicates as expected that the cancer will be better with treatment 1 for stage I ductal white patients. This result leads to the same result we get from the estimate of the parameter $\beta_2$ as we expected.

![Figure 6.6 Stage I Breast Cancer Intensity Function with Treatment 1](image1)

![Figure 6.7 Stage I Breast Cancer Intensity Function with Treatment 2](image2)
The graph given by Figure 6.7, above, shows that as the cumulative time of the patients increase, the intensity function will also increase which indicate as expected that the cancer progress without treatment. This result verifies the result we get from the estimate of the parameter 13. The graph given by Figure 6.8, below, shows that as the cumulative time of the patients increase, the intensity function will also decrease which indicate as expected that the cancer will be better with treatment 1 for stage 1 ductal white patients. This result leads to the same result we get from the estimate of the parameter 14 as we expected.

Figure 6.8 Stage I Breast Cancer Intensity Function with Treatment 3

Following the similar method, the graphs given by Figure 6.9, Figure 6.10 and Figure 6.11, below, show that as the cumulative time of the patients increases, the intensity function will also increase. The intensity function will also increase which indicates as expected that the cancer progresses without treatment or with treatment 1 or 2 for stage II ductal white patients. This result leads to the same result we get from the estimate of the parameter 21, 22 and 23 as we expected.
Figure 6.9 Stage II Breast Cancer Intensity Function without Treatment

Figure 6.10 Stage II Breast Cancer Intensity Function with Treatment 1
Figure 6.11 Stage II Breast Cancer Intensity Function with Treatment 2

The graph given by Figure 4.12, below, shows that as the cumulative time of the patients' increases, the intensity function will also decrease, which indicates as expected that the cancer will be better with treatment 3 for stage II ductal white patients. This result attests to the estimation we get from the estimate of the parameter $\theta_4$ from Figure 6.4.

Figure 6.12 Stage II Breast Cancer Intensity Function with Treatment 3

In summary, we found for white ductal breast cancer patients it is recommended to conduct either combination treatment or beam radiation when they are in early stages like I and II.
6.3 Conclusions

Based on breast cancer patients from SEER database, we have enough information to apply the PLP analysis for white ductal female patients in early two stages. Based on the response of applying proposed model, we can conclude the following.

With no treatment, the intensity function in stage I and stage II increases exponentially which implies that the tumor size of the patients increases at the same rate.

With treatment 1, beam radiation, in stage I the intensity function decreases which implies that the tumor size decrease. However, same treatment in stage II will show opposite result.

With treatment 2, radioactive implants, the intensity function in stage I increases and similar behavior for same treatment in stage II which implies that the tumor size of the patients increase at the same rate.

With treatment 3, combination treatment, the intensity function in stage I and stage II decreases exponentially which implies that the tumor size of the patients decrease at the same rate.

We will continue this study for other races, other types of breast cancer with all four stages. We will continue do the study for more data and eventually we wish to construct PLP for each stage, each tumor size available for all treatment and compare the results. After that we can make suggestion for patients with particular tumor size what kind of treatment is best for them in term of tumor size. We can also apply PLP in Bayesian survival analysis to compare and improve the results.
Chapter 7

Statistical Modeling of Breast Cancer Using Differential Equations

7.1 Background and Data

The object of the present study is to develop differential equations that will characterize the behavior of the tumor as a function of time. Having such differential equations, the solution of which once plotted will identify the rate of change of tumor size as a function of age. The structures of the differential equations characterize the growth of breast cancer tumor. Once we have developed the differential equations and their solutions, we proceed to validate the quality of the differential system and discuss its usefulness.

Breast cancer (malignant breast neoplasm) is cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk, from Sariego (2010). There are different types of breast cancer. The object of the present study is to develop a differential equation that characterizes the behavior of the tumor as a function of age. With respect to the present study, we will address several questions:

What is the mathematical characterization of the growth of the breast cancer tumor as a function of age?

Is the analytical behavior of breast tumor size (TS) uniform over all age?

If the mathematical behavior of the tumor size as a function of age is not uniform, can we identify the age intervals where the sizes of the tumor have the same analytical growth behavior?
Can we identify and justify their mathematical behavior of the size of tumor as a function of age over these age intervals?

Can we develop a differential equation in characterizing the change of breast tumor size as a function of age over these age intervals?

In the present analysis, we used real data that we obtained from Surveillance Epidemiology and End Results (SEER) Program supported by NIH (2010). SEER collects information on incidence, survival, and prevalence from specific geographic areas representing 26 percent of the U.S. population and compiles reports on all of these items plus cancer mortality for the entire U.S.

The proposed differential equations are useful in predicting the rate of change of the cancer tumor size for a specified age of interest. The quality of the differential equation is statistically evaluated using residual analysis. Finally, the usefulness of our findings is discussed.

The proposed differential equation is useful in predicting the rate of change of the cancer tumor size for a specified age of interested. The quality of the differential equation is statistically evaluated using residual analysis. Finally, the usefulness of our findings is discussed.

Many researchers had been working on various aspects of breast cancer. Here, we have summarized some of the most recent and relevant researches to the present study. Madigan et. al. (1995) studied the risk factors for breast cancer patients in the United States. Winchester (1996) presented the relationship between breast cancer and age. Feig and Hendrick (1997) analyzed the risk related to women aged 40 to 49 who undergo mammography procedure. Venturi (2001) mentioned the key role for iodine in breast diseases. Fyles et.al. (2004) studied the behavior of Tamoxifen with or without breast radiation for women 50 years of age or older with early breast cancer detection. Jayasinghe (2005) studied the behavior of age as one factor for the breast cancer patients. Chlebowski et. al. (2006) introduced the
relationship between interim efficacy for female nutrition and cancer. Boffetta et. al. (2006) conducted research on the relation between alcohol drinking and cancer. Ibrahim et.al. (2008) presented a decision tree analysis for competing risks in breast cancer. Buchholz (2009) discussed the benefit of radiation therapy for early-stage breast cancer patients. Xu, Kepner and Tsokos (2011) have introduced a statistical model of breast cancer tumor size that is used to identify the attributable variables and significant interactions and ranking their influences.

7.2 Methodology

From the SEER data we have randomly selected information of 1,000 breast cancer patients. The age of the patients ranges from 33 to 85 years old. However, from the age of 33 to 40 year old patients the data is not complete. Thus, our analysis is focused from the age of 41 to 85 year of age. The scatter diagram given by Figure 2.1, below, is obtained by averaging breast tumor sizes as a function of the age of the breast cancer patients.

For better analytical characterization, we decided that our analysis should be based on partitioning the scatter plot into three age intervals. Age Group I will be partitioned into the age interval from 41 to 58 and Age Group II are III from 59 to 73 and 74 to 85 year old, respectively. The data tree diagram given below identifies the sample sizes and the age intervals, Figure 2.2.

Let X stands for the patients' age in term of years and the according tumor size is function $T(x)$ in term of mille meter (mm) then the instantaneous rate of change (IROC) of tumor size is the derivative of the tumor size function respect to time $(T'(x))$.  

104
Figure 7.1 Breast Cancer Patients' Tumor Size from Age 41 to Age 85

Figure 7.2 Breast Cancer Data Tree Diagram
AGE GROUP I

This group consists of 155 patients from 41 to 58 years of age. The scatter diagram of the data with the best fit is given Figure 7.3, below.

![Graph showing scatter plot with best fit line]

Figure 7.3 Breast Cancer Patients' Tumor Size from Age 41 to Age 58

The mathematical function that characterizes the breast cancer tumor size behavior in the given age group is given by the following polynomial 7.1.

\[
T(x) = 2.016 \times 10^6 + 2.518 \times 10^5 x + 1.307 \times 10^4 x^2 + 3.605 \times 10^2 x^3 + 5.581x^4 + 4.593 \times 10^2 x^5 + 1.571 \times 10^4 x^6, \quad 41 \leq x \leq 58 \tag{7.1}
\]

We check the quality of the fitting by residual analysis of the breast cancer tumor size in the following table 7.1.

Thus, based on the residual analysis we can conclude that the analytical behavior of the tumor size of breast cancer patients given by equation 7.1 is a good fit. The figure 7.4 below shows the actual polynomial over the scatter data.

Now we proceed to identify the differential equation for the first age group. Let \( X \) represent the patients' age in term of years and the according tumor size is a
function, \( T(x) \), in term of millimeter (mm) then the instantaneous rate of change (IROC) of tumor size is the derivative of the tumor size function with respect to time (\( T'(x) \)).

Table 7.1 Residual Analysis of Breast Cancer Tumor Size in Group I

<table>
<thead>
<tr>
<th>Age</th>
<th>Actual value</th>
<th>Fitted value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>19</td>
<td>18.5811</td>
<td>.4189</td>
</tr>
<tr>
<td>42</td>
<td>15</td>
<td>17.824</td>
<td>-2.824</td>
</tr>
<tr>
<td>43</td>
<td>25</td>
<td>19.0416</td>
<td>5.9584</td>
</tr>
<tr>
<td>44</td>
<td>14.75</td>
<td>19.9738</td>
<td>5.2238</td>
</tr>
<tr>
<td>45</td>
<td>23.33</td>
<td>19.856</td>
<td>3.4773</td>
</tr>
<tr>
<td>46</td>
<td>16.25</td>
<td>18.8271</td>
<td>-2.577</td>
</tr>
<tr>
<td>47</td>
<td>16</td>
<td>17.4501</td>
<td>-1.45</td>
</tr>
<tr>
<td>48</td>
<td>16.8</td>
<td>16.3467</td>
<td>.4533</td>
</tr>
<tr>
<td>49</td>
<td>18.75</td>
<td>15.9438</td>
<td>2.8062</td>
</tr>
<tr>
<td>50</td>
<td>18.33</td>
<td>16.3342</td>
<td>1.9991</td>
</tr>
<tr>
<td>51</td>
<td>17.125</td>
<td>17.2499</td>
<td>-.1249</td>
</tr>
</tbody>
</table>

Mean of Residual 7.31957e-17
Standard Deviation of Residual (SD) 3.975954
Standard Error of Residual (SE) 0.9371414

Figure 7.4 Breast Cancer Patients' Tumor Size
from Age 41 to Age 58 with Curve Fitting

The differential equation is given by 7.2, below.

\[
T(x) + T'(x) = 1.764 \times 10^6 \times 2.2456 \times 10^5 x + 1.198 \times 10^4 x^2
3.38 \times 10^2 x^3 + 5.351 x^4 + 0.499 \times 10^2 x^5 + 1.571 \times 10^4 x^6, 41 \times 58 \tag{7.2}
\]

Thus, the solution of the above differential equation 7.2 is given by 7.3, below.

Therefore, if one is interest in obtaining the change of rate of the breast cancer tumor size for a desired age in age group I, he can evaluate the solution of the differential equation at the desired age.

\[
\frac{dT(x)}{dx} = 2.518 \times 10^5 + 2.613 \times 10^4 x + 1.082 \times 10^3 x^2
+ 22.3221 \times 3 \times 0.2297 x^4 + 0.942 \times 10^3 x^5, 41 \times 58 \tag{7.3}
\]

The graph 7.5 given below is a representation of the solution of the differential equation in 7.3. We proceed to evaluate the results given by the solution to the differential equation. We evaluate the accuracy of the results from the differential equation as follows. For example, at age of 41 to 42, the solution to the differential equation estimate the change of rate is -0.04074, where the observed actual rate of change is given by -0.210526 which is obtained from \( ROC = \frac{\text{current year-previous year}}{\text{previous year}} \).

The difference of the two constitutes the first rate of change residual (ROC residual).

The table 7.2 below gives the 10 estimates of the solution of the differential equation.

Based on the below results, we can conclude that the differential equation gives fairly accurate rate of the change of the breast tumor size as a function of age.

We can utilize the mathematical expression shown in equation 7.2 and 7.3 with the correction factor of the mean of residual to estimate the rate of the tumor growth for future age.
Figure 7.5 Breast Cancer Patients’ Tumor Size IROC from Age 41 to Age 58

Table 7.2 Residual Analysis of ROC of Breast Cancer Tumor Size in Group I

<table>
<thead>
<tr>
<th>Age</th>
<th>Empirical ROC</th>
<th>DE IROC</th>
<th>ROC Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>-0.210526</td>
<td>-0.04074</td>
<td>-0.16978212</td>
</tr>
<tr>
<td>43</td>
<td>0.666667</td>
<td>0.0683115</td>
<td>0.598355</td>
</tr>
<tr>
<td>44</td>
<td>-0.410000</td>
<td>0.0489545</td>
<td>-0.458955</td>
</tr>
<tr>
<td>45</td>
<td>0.581921</td>
<td>-0.0058972</td>
<td>0.587818</td>
</tr>
<tr>
<td>46</td>
<td>-0.303571</td>
<td>-0.0518198</td>
<td>-0.251752</td>
</tr>
<tr>
<td>47</td>
<td>-0.015385</td>
<td>-0.0731362</td>
<td>0.057752</td>
</tr>
<tr>
<td>48</td>
<td>0.050000</td>
<td>-0.0632339</td>
<td>0.113234</td>
</tr>
<tr>
<td>49</td>
<td>0.116071</td>
<td>-0.0246478</td>
<td>0.140719</td>
</tr>
<tr>
<td>50</td>
<td>-0.022222</td>
<td>0.024487</td>
<td>-0.046709</td>
</tr>
<tr>
<td>51</td>
<td>-0.065909</td>
<td>0.0560584</td>
<td>-0.121967</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of Residual</td>
<td>0.1012375</td>
</tr>
<tr>
<td>Standard Deviation of</td>
<td>0.4946078</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
</tr>
<tr>
<td>Standard Error of Resi</td>
<td>0.1165802</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AGE GROUP II

This group consists of 276 patients from 59 to 73 years of age. The scatter diagram of the data with the best fit is given Figure 7.6, below.

Figure 7.6 Breast Cancer Patients’ Tumor Size from Age 59 to Age 73

The mathematical function that characterizes the breast cancer tumor size behavior in the given age group is given by the following polynomial 7.4.

\[ T(x) = 1.749 \times 10^5 + 1.0788 \times 10^4 x + 2.491 \times 10^2 x^2 + 2.551x^3 + 9.776 \times 10^3 x^4 \]  

(7.4)

We check the quality of the fitting by residual analysis of the breast cancer tumor size in the following table 7.3.

Thus, based on the residual analysis we can conclude that the analytical behavior of the tumor size of breast cancer patients given by equation 7.4 is a good fit. The figure 7.7 below shows the actual polynomial over the scatter data.
Table 7.3 Residual Analysis of Breast Cancer Tumor Size in Group II

<table>
<thead>
<tr>
<th>Age</th>
<th>Actual value</th>
<th>Fitted value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>18.875</td>
<td>18.81482</td>
<td>0.06018</td>
</tr>
<tr>
<td>59</td>
<td>18.8235</td>
<td>16.72068</td>
<td>2.10285</td>
</tr>
<tr>
<td>60</td>
<td>13.2381</td>
<td>16.88538</td>
<td>-3.64728</td>
</tr>
<tr>
<td>61</td>
<td>16.2105</td>
<td>18.2</td>
<td>-1.989647</td>
</tr>
<tr>
<td>62</td>
<td>23.3333</td>
<td>19.79</td>
<td>3.5423886</td>
</tr>
<tr>
<td>63</td>
<td>20.8</td>
<td>21.018</td>
<td>-0.21818</td>
</tr>
</tbody>
</table>

Mean of Residual 1.484996e-17
Standard Deviation of Residual 2.706574
Standard Error of Residual 0.6988345

Figure 7.7 Breast Cancer Patients’ Tumor Size
from Age 59 to Age 73 with Curve Fitting

Now we proceed to identify the differential equation for the second age group.

The differential equation is given by 7.5, below.

\[
T(x) + T' (x) = 1.6414 \times 10^5 + 1.029 \times 10^4 x + 2.4141 \times 10^2 x^2 + 2.511 \times 10^3 x^3 + 9.776 \times 10^3 x^4 , \quad 59 < x < 73 \tag{7.5}
\]

The instantaneous rate of change of breast cancer patients’ tumor size as a function of time is given analytically by

111
\[
\frac{d(T(x))}{d(x)} = 1.0788 \times 10^4 + 4.9812 \times 10^2 x \\
7.652 x^2 + .0391 x^3, \quad 59 \times 73
\] (7.6)

A graphical display of expression above is given by Figure 2.6, below.

![Graph of Tumor Size IROC](image)

Figure 7.6 Breast Cancer Patients’ Tumor Size IROC from Age 59 to Age 73

The residual analysis we performed on the proposed differential equation of tumor size is given in Table 7.4, below. We will only keep 5 residual data.

Table 7.4 Residual Analysis of ROC of Breast Cancer Tumor Size in Group II

<table>
<thead>
<tr>
<th>Age</th>
<th>Empirical ROC</th>
<th>DE IROC</th>
<th>ROC Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>-0.4529</td>
<td>-0.40292</td>
<td>-0.04998</td>
</tr>
<tr>
<td>60</td>
<td>-0.00273</td>
<td>-0.2513</td>
<td>0.24857</td>
</tr>
<tr>
<td>61</td>
<td>-0.29673</td>
<td>0.005681</td>
<td>-0.30241</td>
</tr>
<tr>
<td>62</td>
<td>0.224536</td>
<td>0.164981</td>
<td>0.059555</td>
</tr>
<tr>
<td>63</td>
<td>0.439394</td>
<td>0.161232</td>
<td>0.278162</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean of Residual</th>
<th>0.01485848</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Deviation of Residual</td>
<td>0.1969682</td>
</tr>
<tr>
<td>Standard Error of Residual</td>
<td>0.05462916</td>
</tr>
</tbody>
</table>

As seen from the table above the residuals are small and so is the standard error. These results attest to the good quality of the proposed model for tumor size
AGE GROUP III

This group consists of 308 patients from 73 to 85 years of age. The scatter diagram of the data with the best fit is given Figure 7.9, below.

![Figure 7.9 Breast Cancer Patients' Tumor Size from Age 73 to Age 85](image)

The mathematical function that characterizes the breast cancer tumor size behavior in the given age group is given by the following polynomial 7.7.

\[
T(x) = 2.93789 \times 10^5 + 1.4954 \times 10^4 x + 2.853 \times 10^2 x^2 + 2.4166 x^3 + 7.672 \times 10^3 x^4, \quad 74 \leq x \leq 85
\] (7.7)

We check the quality of the fitting by residual analysis of the breast cancer tumor size in the following table 7.5.

Thus, based on the residual analysis we can conclude that the analytical behavior of the tumor size of breast cancer patients given by equation 7.7 is a good fit. The figure 7.10 below shows the actual polynomial over the scatter data.
Table 7.5 Residual Analysis of Breast Cancer Tumor Size in Group III

<table>
<thead>
<tr>
<th>Age</th>
<th>Actual value</th>
<th>Fitted value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>15.16667</td>
<td>14.53395</td>
<td>0.63271232</td>
</tr>
<tr>
<td>75</td>
<td>15.26667</td>
<td>16.05426</td>
<td>-0.78759789</td>
</tr>
<tr>
<td>76</td>
<td>15.48387</td>
<td>16.69901</td>
<td>-1.2151395</td>
</tr>
<tr>
<td>77</td>
<td>17.9</td>
<td>17.06703</td>
<td>0.83296694</td>
</tr>
<tr>
<td>78</td>
<td>19.11765</td>
<td>17.57305</td>
<td>1.5445921</td>
</tr>
<tr>
<td>79</td>
<td>18.45833</td>
<td>18.44768</td>
<td>0.01065</td>
</tr>
</tbody>
</table>

Mean of Residual: -6.780781e-18
Standard Deviation of Residual (SD): 0.929734
Standard Error of Residual (SE): 0.2683911

Figure 7.10 Breast Cancer Patients’ Tumor Size from Age 74 to Age 85 with Curve Fitting

Now we proceed to identify the differential equation for the third age group.

The differential equation is given by 7.8, below.

\[
T(x) + T'(x) = 2.7883 \times 10^7 + 1.438 \times 10^4 x + 2.78 \times 10^2 x^2 + 2.3859 \times 7.672 \times x^{3/4} \times 74 \times 85
\]  

(7.8)

Thus, the solution of the above differential equation 7.8 is given by 7.9, below.

Therefore, if one is interested in obtaining the change of rate of the breast cancer
tumor size for a desired age in age group III, he can evaluate the solution of the differential equation at the desired age.

\[
\frac{dT(x)}{dx} = 1.4954 \times 10^4 \times 5.705 \times 10^2 \times x^2 + 7.24988 \times 0.03068 \times x^3, \quad 74 \leq x \leq 85
\]

(7.9)

A graphical display of expression above is given by Figure 7.11, below.

Figure 7.11 Breast Cancer Patients’ Tumor Size IROC from Age 73 to Age 85

The residual analysis we performed on the proposed differential equation of tumor size is given in Table 7.6, below. We will only keep 5 residual data.

Table 7.6 Residual Analysis of ROC of Breast Cancer Tumor Size in Group III

<table>
<thead>
<tr>
<th>Age</th>
<th>Empirical ROC</th>
<th>DE IROC</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>0.0065934</td>
<td>-0.030347</td>
<td>0.036941</td>
</tr>
<tr>
<td>75</td>
<td>0.014227</td>
<td>0.0967895</td>
<td>-0.082562</td>
</tr>
<tr>
<td>76</td>
<td>0.156042</td>
<td>0.090693</td>
<td>0.06534849</td>
</tr>
<tr>
<td>77</td>
<td>0.06802498</td>
<td>0.04762</td>
<td>0.0204083</td>
</tr>
<tr>
<td>78</td>
<td>-0.0344872</td>
<td>0.0187968</td>
<td>-0.053284</td>
</tr>
</tbody>
</table>

Mean of Residual | 0.001199553
Standard Deviation of Residual (SD) | 0.04160096
Standard Error of Residual (SE) | 0.01200916
As seen from the table above the residuals are small and so is the standard error. These results attest to the good quality of the proposed model for tumor size. We can conclude from our extensive statistical analysis that all of the four parts of the differential equations have good quality.

This model is useful for a number of reasons.

1. It can be used to identify the rate of change of the growth of the breast cancer tumor size.

2. One can also use the proposed differential equation systems to generate various scenarios of the tumor size as a function of different values of the age.

3. People can use these differential equation systems to predict the rate of change of the growth of tumor for different ages.

7.3 Conclusions

In the present study, we extract a random sample 1,000 breast cancer patients from SEER data base and develop differential equations to obtain information about the rate of growth of breast cancer tumor. We found the breast cancer tumor size is not uniform over all age. The sample data was partitioned into three intervals groups as a function of age for better analytical tractability, that is, the age group from 41 to 58, age group from 59 to 73 and age group from 74 to 85. For each age group, we develop a differential equation that can be used to obtain the rate of growth of the malignant tumor size. We justified the mathematical behavior of the function we proposed by residual analysis.
Chapter 8
Conclusions and Future Research

8.1 Conclusions

In the present study we have developed the actual differential equations that characterize the behavior of each of the six attributable variables that constitute the carbon dioxide emissions, namely, gas fuels, liquid fuels, solid fuels, flaring, cement, and bunker. We have developed the differential system of the sum of the six attributable variables that constitute CO2 emissions using actual data from 1950 to 2005 for continental United States. In addition to have given the analytical form for each variable, we have used three different statistical procedures, namely $R^2$ ($R^2$ adjusted), PRESS statistic and residual analysis to evaluate the quality of the proposed differential methods. All these statistical procedures attest to the quality of the proposed differential systems. Finally, we have used these models to make 10, 20, and 50 year prediction to the rate of change of the entities that constitute CO2 emissions. This information should be useful for strategic planning and formulating policies to assist in the problem of GLOBAL WARMING.

We have developed the actual differential equations that characterize the behavior of each of the five attributable variables that constitute the carbon dioxide in the atmosphere, namely, gas fuels, liquid fuels, solid fuels, flaring, cement, and bunker. We have developed the differential system of the total CO2 in the atmospheres using actual data from 1959 to 2004 for the continental United States. In addition to having given the analytical form for each variable, we have used three different statistical procedures, namely $R^2$ ($R^2$ adjusted), PRESS statistic and residual analysis to evaluate the quality of the proposed differential methods. All
these statistical procedures attest to the quality of the proposed differential systems. Finally, we have used these models to make 10, 20, and 50 year predictions to the rate of change of the entities that constitute CO2 in the atmosphere. This information should be useful for strategic planning and formulating policies to assist in the problem of GLOBAL WARMING.

We have performed parametric analysis for CO2 in the atmosphere. The initial measurement of CO2 in the atmosphere was collected at Mauna Loa Observatory, Hawaii (C.D. Keeling, T.P. Whorf, 2005). Those data do not follow normal probability distribution. Thus, we transform the response of the data by using Box-Cox transformation that resulted in make the CO2 being normal. We proceed to develop a “nonlinear” statistical model (nonlinear in terms of the higher power of the response variable). Through the process of developing the statistical model, we have found that only five variables, namely, Liquid, Bunker, Cement, Gas flares and Gas fuels significantly contribute to the CO2 in the atmosphere with five interactions among them. The proposed statistical model was evaluated using the R-square, R-square adjusted and PRESS statistics. We also provide model cross validation by hiding some part of the data and estimate them from the rest of the data. The mean residuals of those cross validation are extremely small. All criteria results support the high quality of the developed statistical model. Furthermore we have ranked the individual attributable variable and interaction according to their contribution to CO2 in the atmosphere.

We can conclude from our extensive statistical analysis that there are only five significant attributable variables to the CO2 in the atmosphere namely, Gas fuels, Gas flares, Bunker, Liquid and Cement. Furthermore, we also tested all possible 2nd order interactions of all attributable variables and we found only five interactions that significantly contribute to CO2 in the atmosphere, namely, Liquid with Cement, Cement with Bunker, Gas Fuels with Gas Flares, Liquid with Gas Flares and Gas
Flares with Bunker. Thus, one may obtain a good estimate of the CO2 in the atmosphere by knowing the measurement of Cement and those five interactions.

One can utilize the above model equation to perform surface response analysis to identify the values of the contributable variables that will minimize the CO2 in the atmosphere.

For cancer research, we can conclude from our extensive statistical analysis that there are only three significant contributable variables to the tumor size for breast cancer namely, R XR(X14), RXPS(X13) and AJCC(X19). As for SNC(X5), HT(X10). They themselves individually do not significantly contribute to the response variables; however, when they interact with other variables, they do significantly contribute to the response variable. Furthermore, we also tested two thousand and three hundred possible interactions of the contributable variables and we found three interactions to significantly contribute to tumor size for breast cancer.

This model is useful for a number of reasons.

1. One can also use the proposed model to generate various scenarios of the breast cancer tumor size as a function of different values of the subjective entities for data simulation purpose.

2. It can be used to identify the significant contributable variables.

3. It identifies the significant interactions of these contributable variables.

4. The significant contributions to the breast cancer tumor size are ranked.

5. A confidence interval for the tumor size can be constructed with parametric analysis. By obtaining the 100(1 - %) confidence limits for the response, we can describe how confident we are that our estimate is close to the actual tumor size.

6. The model as shown in equation 3 can be used to perform surface response analysis to place the restrictions on the significant contributable variables and interactions to minimize the breast cancer tumor size. We can also put
restrictions on the variables to minimize the response of the tumor size by nonlinear
control with 100(1)\% confidence limits.

We performed parametric analysis to estimate tumor size for breast cancer
patients for data simulation purpose. The initial measurement of tumor size was
collected from the SEER database. Those data do not follow normal probability
distribution. Using the standard Box-Cox transformation, the SEER tumor size data
became approximately normally distributed. We developed a “nonlinear” statistical
model (nonlinear in terms of the power and logarithm of the response variable).
Through the process of developing the statistical model, we found only four variables,
namely, rxx(X14), rxps(X13) and ajcc(X19) and three interactions that significantly
contribute to the tumor size. The proposed statistical model was evaluated using the
R-square, R-square adjusted, PRESS statistics and three cross validation methods,
all of which support the high quality of the developed statistical model. This model
can be used to obtain a good estimate of tumor size knowing the four significantly
attributable variables and three interaction terms.

We validate this model on four different data sets. Since this type of model is
strongly data driven, we need to make the necessary modification. For example, the
Box-Cox transformation’s power estimation needs to be changed according to the
data set. The specific model’s interaction and estimation of the coefficients need to
be recalculated. After the necessary modifications these four data sets fit the model
as following. The model fits one of the data sets pretty well with high R-square more
than 90. The model fit two data sets not very well with R-square around 30 and one
data set pretty bad with R-square around 10. Considering the complexity and random
behavior of the breast cancer and the limitation of our available data set, we are not
surprised with the result. We wish to make the update of the model based on large
scale data set for the future study.
8.2 Future Research

8.2.1 Global Warming

In Global warming the air temperature plays an important role. I plan to do some research study so that people can predict the CO2 in atmosphere by obtaining the air temperature. We can conduct more study on this area.

8.2.2 Cancer Research

We can develop some prescreening process with high confidence level by which people can predict the odds of the cancer by develop an early risk detection model. I have begun this wok with Dr. Charles E. Cox from USF College of Medicine. We are revising so called Gail model to predict the risk of the breast cancer. Eventually I hope we can use this modified risk model to provide a prescreening model for many different cancers. As for the evaluation of the cancer treatments, I have constructed several PLP models and I have verified these models can provide evaluations of different treatments for stage I and stage II breast cancer patients. I hope in further research we can improve this model so that it contains more factors such as age, genotype and we can handle censored information with PLP. Also, we will modify these PLP models for all kinds of cancers. We can continue do the study for more data and eventually we can construct PLP for each stage, each tumor size available for all treatment and compare the results. Then we can make suggestion for patients with particular tumor size that which treatment is best for them to have maximum survival expectation. We will fix grade and behavior code with same sex, stage and tumor size. We can expand this form uncensored case to censored case. We can also apply PLP in Bayesian survival analysis to improve our result and give better suggestions.
References


Fyles, A. W., McCready, D. R., Manchul, L. A. and et al., 2004. Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer. New England Journal of Medicine, 351, 963-970.


Habibi, G., S. Leung, J.H. Law, K. Gelmon and H. Masoudi et al., 2008. Redefining prognostic factors for breast cancer: YB-1 is a stronger predictor of
relapse and disease-specific survival than estrogen receptor or HER-2 across all tumor subtypes. BCR., 10: 86.


Winchester, D. P., 1996. The National Cancer Data Base report on breast carcinoma characteristics and outcome in relation to age. CANCER 78 : 1838


About the Author

Yong Xu got his bachelor degree in EE at Shandong University in China 2001. When he worked as quality control engineer after graduation, he had strong interest in statistics and decided to pursue statistics for future career.

He got his master degree in statistics at University of North Florida at fall, 2004. He taught statistic and mathematics classes as adjunct professor at UNF for one year in 2005. He worked as assistant professor in the department of mathematics and statistics at Shandong Economic University, China in 2006.

He continued his graduate level study at University of South Florida for Ph.D. degree with a concentration in Statistics in fall 2006. He got summer internship at Statistical Evaluation Center in American Cancer Society in 2009 and he worked as research intern for Moffitt Cancer Center Biostatistics Core in 2010. His main research interests are Biostatistics and Statistics for Environmental Sciences.