November 2017

Statistical Analysis and Modeling of Stomach Cancer Data

Chao Gao
University of South Florida, cgao@mail.usf.edu

Statistical Analysis and Modeling of Stomach Cancer Data

By

Chao Gao

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

Major Professor: Chris P. Tsokos, Ph.D.
Lu Lu, Ph.D.
Kandethody Ramachandran, Ph.D.
Dan Shen, Ph.D.

Date of Approval:
November 14, 2017

Keywords: stomach cancer, parametric analysis, tumor size,
quantile regression, survival analysis

DEDICATION

I would like to dedicate my dissertation to my wife, my daughter, my parents, for all of their support and encouragement.

I would also want to dedicate my dissertation to my advisor Dr. Chris Tsokos. He always teaches me and support me during last several years.

ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Chris Tsokos, for his great help and suggestions throughout my graduate studies at the University of South Florida. I was truly enjoyed the wonderful relationship with Dr. Tsokos. He guided me understand the importance of statistical modeling with the extensive background in many application fields. He is a very knowledge person and a good friend.

I would also want to thank my dissertation committee for their great efforts and useful advice throughout my dissertation process.

Finally, I would like to thank my wife, my parents, and my daughter for their unconditionally support and encouragement.

Table of Contents

List of Tables

List of Figures

Abstract

The objective of this study is to address some important questions associated with stomach cancer patients using the data from the Surveillance Epidemiology and End Results (SEER) program of the United States. To better understand the behavior of stomach cancer, we first perform parametric analysis for each patient group (white male, white female, African American male, African American female, other male and female) to identify the probability distribution function which can best characterize the behavior of the malignant stomach tumor sizes. We evaluate the effects of patients' age, gender and race on the malignant stomach tumor sizes by using quantile regression models, which gives us a better understanding of the behavior of the malignant stomach tumors.

We also developed statistical models with respect to patients' malignant stomach tumor size as a function of age for different races and gender group, respectively. The developed models were evaluated to attest their prediction quality. Furthermore, we have identified the rate of change of the malignant tumor size as a function of age, for gender and race.

We evaluated the routine treatment of stomach cancer using parametric and nonparametric survival analysis. We have found that stomach cancer patients who receive surgery with radiation together have a better survival probability than the patients who receive only radiation.

We performed decision tree analysis to assist the physician in recommending to his patients the most effective treatment that is a function of their characteristics.

Chapter 1

Introduction

1.1 Cancer

In medical research, the tumor refers to the cells' changing abnormality that is not necessarily meaning the body tumor. Some body parts of the cells will result in uncontrolled proliferation by the disease (www.wikipedia.org). In 2017 in USA, there will be an estimated 1,688,780 new cancer cases and 600,920 cancer deaths. That is approximately 4,630 new cases and 1,650 deaths per day. From 1991 to 2014, the total cancer death rate was reduced 25% than the expected (www.cancer.org). Such progress benefitted from treatment improvements and earlier diagnoses.

1.2 Stomach Cancer

According to the American Cancer Society, stomach cancer is the fifth leading cause of cancer and third leading cause of death from stomach cancer. Stomach cancer, also called gastric cancer, is a cancer that starts in stomach. Some people like to use "stomach" to mention the body parts between the pelvic and the chest. The medical term for this area is called "Abdomen". Stomach cancer should not be puzzled with other cancers that can exist in the abdomen, like colon cancer, liver cancer because those different cancers have different treatments, different symptoms, and different outlooks.

In 2017, there are about 28,000 cases of stomach cancer will be diagnosed (17,750 men and 10250 women) and about 10,960 people will die from this type of cancer (6,720 men and 4240 women) that was estimated by American Cancer Society. If the physicians find the stomach tumor while only in the stomach, the 5-year survival rate is about 65%. If they find the stomach tumor spread to areas near your stomach or lymph nodes, the 5-year survival rate is about 30%. If they find the stomach cancer spread far away from the stomach, the 5-year survival rate will be reduced to about 5%. Thus, the overall 5-year survival rate of all people with stomach cancer in the United States is about 29%. The 5-year relative survival probability compares the observed survival of patient with stomach cancer to what is expected for the person without stomach cancer. Since some people may die from other disease or other types of cancer, this is a better way to see the impact of cancer on survival. This survival rate has improved over the past 30 years. The most important reason that the total survival probability is pretty low in the United States is that most of stomach cancer patients are diagnosed at a later stage. The stage of the cancer plays a major role of the survival rate. The stage of a cancer is a description of how far the cancer has extend. The stomach cancer's stage is very important for the physicians to choose the best treatment for their patients. There are two types of stages for stomach cancer patients. The clinical stage of the stomach cancer is the best time for the doctors to treat the cancer of their patients based on the results of physical exams, endoscopy, biopsies, and any imaging tests such as CT scans, etc. Once the surgery is done, the pathologic stage can be determined using the same results from the clinical stage.

A risk factor will affect patients' chance of getting a cancer. Different cancers always have different risk factors. There are several risk factors for stomach cancer. As shown in Figure 1.1, Gender, age, Ethnicity, Geography, Helicobacter pylori infection, Tobacco use, diet seem to play a role in raising the risk of a stomach cancer.

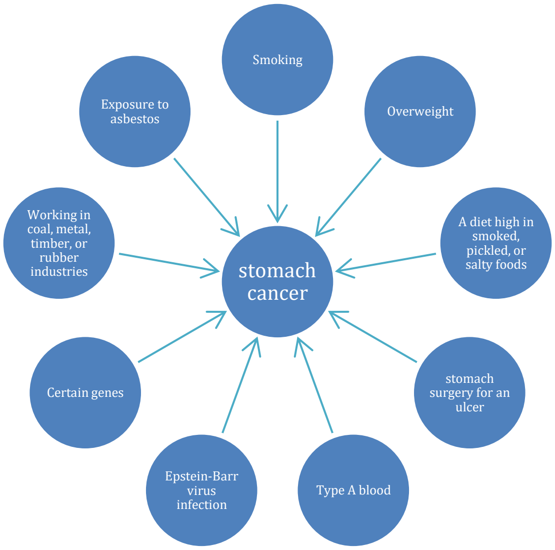

Smoking
Exposure to asbestos
Overweight
Working in coal, metal, timber, or rubber industries
stomach cancer
A diet high in smoked, pickled, or salty foods
Certain genes
stomach surgery for an ulcer
Epstein-Barr virus infection
Type A blood

Figure 1.1: potential risk factors of stomach cancer

Stomach cancer is a very common cancer among East Asian people, especially in Japan. Averagely, the stomach cancer incidence is about 60 per 100,000 people. In 2007, Japan Cancer Society reports that one third of deaths were related to stomach cancer. Almost 70 to 90 percent of all stomach cancers begin with Helicobacter pylori, or H. pylori infection. The H. pylori bacteria spread by unwashed or undercooked foods. The

Japanese diet is characteristically high in seasoned foods. Salted food consumption is one of the most important causes of increased stomach cancer risk. In Japan, there about 30 out of 100,000 people will develop stomach cancer in their whole life. The Japanese government encourages their people to take screen test for stomach every year once they older than age 40. That is the best way to find stomach cancer and detect it during the earliest stage.

There is no sure way to prevent stomach cancer, but as the stomach tumors grow, the patients may have more serious symptoms as shown in Figure 1.2 below. The best treatment options for stomach cancer patients are Surgery, Chemotherapy, Targeted Therapy, and Radiation Therapy. The doctors often suggest their patients to use two or more of those treatment methods. But they do not know which treatment is the best for their patients. Thus, we will also want to weight the benefits from each different treatment and their potential side effects. The treatment selection depends on many factors. The location of the tumor and the stage are very important for the doctors and patients to consider. The physicians would also take the patients' age, gender, body situation, and other factors into account.

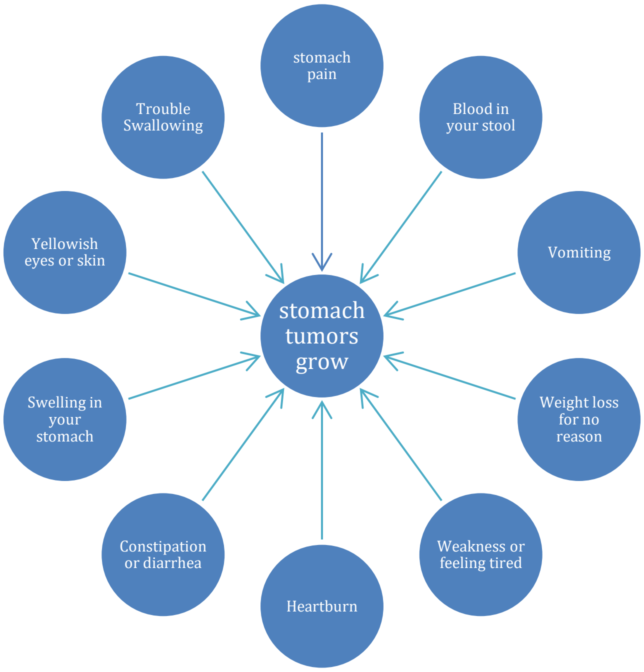


Figure 1.2: more serious symptoms as stomach tumors grow

Researchers have reported parametric models that are similar to the Cox regression (Moghimi-Dehkordi, Bijan, et al. 2008) method. Patients aged 60-75 and >75 years at diagnosis had an increased risk for death. Although the hazard ratio in the Cox model and other parametric models are approximately similar from the AIC criteria. The Exponential and Weibull probability distributions are the most favorable distributions for determine survival analysis. The survival probability and the risk of stomach cancer death among the patients with screen-detected cancer and patients with interval cancer were not significantly different in the annual endoscopic screening (Hamashima, Chisato, et al. 2015). These results suggest that the potential of endoscopic screening in reducing mortality from gastric cancer. Missing value imputation is increasing the estimate of precision and accuracy (Moghimbeigi, Abbas, et al. 2014). The survival rate in Japan was clearly higher than those in the other countries (Matsuda, Tomohiro, and Kumiko

Saika, 2013). The high survival rate for Japanese patients could be related to the stomach cancer screening and abundant experience in treatment according to the high incidence rate. Patients who survive from gastric and gastroesophageal junction more than 3 years after diagnosis have demographic and pathologic characteristics distinct from those who do not survive (Kunz, Pamela L., et al. 2012). The survival of stomach cancer patient diagnosis appears to be increasing (Hansson, Lars-Erik, Pär Sparén, and Olof Nyrén, 1999). The reasons for this are probably multifactorial and are likely to include improvements in anesthesiologic and surgical management. Research scientists have shown that patients' age, gender, and tumor location are significantly independent prognostic factors for overall survival in patients with metastatic gastric cancer (Yang, Dongyun, et al. 2011).

1.3 Research Data

In 1973, the National Cancer Institute funded the Surveillance, Epidemiology, and End Results program (SEER). The SEER program provides national leadership in the health science of cancer surveillance as well as analytical tools and methodological expertise in interpreting, analyzing, collecting, and disseminating reliable population-based statistics. The SEER database is a premier source for various types of cancers in the United States. The SEER program collects population-based cancer registries from 20 different locations across the United States which are Alaska Native Tumor Registry, Arizona Indians, Cherokee Nation, Connecticut, Detroit, Atlanta, Greater Georgia, Rural Georgia, San Francisco-Oakland, San Jose-Monterey, Greater California, Hawaii, Iowa, Kentucky, Los Angeles, Louisiana, New Jersey, New Mexico, Seattle-Puget Sound and Utah. The information about cancer registries covers 28% of the population in the United States by collecting complete and accurate data on all kinds of cancers that have been diagnosed. The SEER cancer statistics review, which is a report of the most recent cancer incidence, mortality, survival, prevalence, and lifetime risk statistics, are published by the Surveillance research program of National Cancer Institute. In the present study, there is a total of 11,462 records of stomach cancer patients from the year 2004 to 2013 were obtained from the SEER program. There are 7,149 males (7,115 malignant) and 4,313 females (4,279 malignant) stomach patients. The probability of tumor in stomach not to be malignant is 0.0048 for male and 0.0078 for female. Our research will focus on the malignant stomach patients.

1.4 Parametric Analysis

The basic idea of parametric analysis of a given set of data is to identify, if possible, the probability distribution function that characterizes the phenomenon that is represented by the given data. Based on the specific probability distribution, we can obtain the important approximate estimates from the parameters such as expected mean, standard deviation, and confidence limits, etc. In Chapter 2, our objective is to find if there is a well-known probability distribution that can be used effectively to characterize the behavior of the stomach cancer patients' tumor size. We first test if there is a significant difference of the average malignant tumor size between the patients' gender and race. We found that the average malignant tumor size is different by patients' gender and race, respectively. In addition, we utilized three goodness of fit tests, Kolmogorove-Smirnov, Anderson-Darling, Chi-Square, to identify the best probability distribution function which can characterize the stomach patients' malignant tumor size. We found that expected average stomach tumor size of males is higher than that of females and African American patients have the largest expected average stomach tumor sizes. The useful information such as expected malignant tumor size along with its variance and confidence limits will assist physicians to make appropriate decisions with a given degree of assurance. Furthermore, there is an exponential growing behavior between the size of a tumor and the probability of the tumor being malignant. When the tumor size is 10 millimeters or above, it has a 99.6% or higher chance that the tumor is malignant. Such findings and the graphical figures of the probability density functions and the cumulative distribution curves could provide assistance for physicians to understand the probabilistic behavior of the stomach tumor size.

1.5 Quantile Regression Model

The average malignant stomach tumor size is often insufficient to explain the probabilistic behavior of malignant stomach tumor size. Standard least square estimates only provide a summary of the risk factors on the average malignant stomach tumor sizes, which may hide important elements of the underlying relationships. In Chapter 2, we found the best probability distributions are highly skewed Wakeby probability distribution, three parameters Weibull probability distribution and Dagum probability distribution, which means the standard least squares assumption of normally distributed errors fails to hold. The quantile regression can exhibit a comprehensive picture when both upper and lower quantiles of the distributions of response are our interests. In Chapter 3, we developed a statistical quantile regression model to describe the effects of patients' race, gender, and age on the sizes of malignant stomach tumors. Based on the results of our statistical quantile regression model, we have identified there is a significant difference between male and female patients. For instance, for the 60th quantile, the malignant tumor size of male patients is 7.89 millimeters larger than females when the other covariates hold the same level. Such information is important to medical doctors to address the procedural and clinical strategies for their patients. Moreover, our statistical quantile model reveals a significant difference of stomach tumor size between patients' race. For instance, for the 45th quantile, African American patients' malignant tumor size is 9.49 millimeter bigger than other race patients'. In addition, we found the patients' age also is a significantly contribution factor for estimating the malignant tumor size. For the 25th quantile, the malignant tumor size increases 0.125 millimeters for one-year increase in patients' age. Our developed statistical quantile regression model gives a more comprehensive comparison of malignant tumor size. Such findings are extremely important for medical doctors to improve their treatments for their patients.

1.6 Statistical Modeling

Statistical modeling is a simplified, mathematically formalized way to approximate reality and optionally to make predictions from this approximation. Under a set of assumptions, we develop the statistical model based on the sample data. A good statistical model can be used to identify the attributable variables to the response variable and to identify the significant interactions among those explanatory variables which contributes to the response significantly. Once we have a good statistical model, we can use it for prediction and forecasting. In addition, we can do surface response analysis. An easy way to estimate a first-degree polynomial model is to use a factorial experiment or a fractional factorial design. This is sufficient to determine which explanatory variables affect the response variable(s) of interest. Once it is suspected that only significant explanatory variables are left, then a more complicated design, such as a central composite design, can be implemented to estimate a second-degree polynomial model, which is still only an approximation at best. However, the second-degree model can be used to optimize (maximize, minimize, etc). Furthermore, a good statistical model can be used to identify how the responses vary for different categories. Moreover, a good statistical model can be used to investigate the relationships between explanatory variables and the response variable. The main objective of Chapter 4 is to identify the stomach cancer tumor size as a function of patient's age. The statistical analysis research was performed under different race and gender groups, respectively. We developed a nonlinear statistical model to fit the observed data and the residual analysis was performed to help us identify the appropriate fit. The differential equation guides us to figure out how the rate of changing mean size of malignant tumor when the patients increase the age.

1.7 Decision Tree Analysis

Decision making always plays a very important role in many fields of research, especially in cancer research. The important idea of survival tree analysis is to split the given sample into subgroups by continuous spiting of the initial node into child nodes based on the uniformity of within-node instances or separation of between-node instances with respect to our risk factors. For each node, risk factors are tested based on the splitting criterion. After the node is divided into two child nodes according to the value of the attribute variable, we repeated the process for each child node. The whole process of a decision tree is shown below in Figure 1.3.

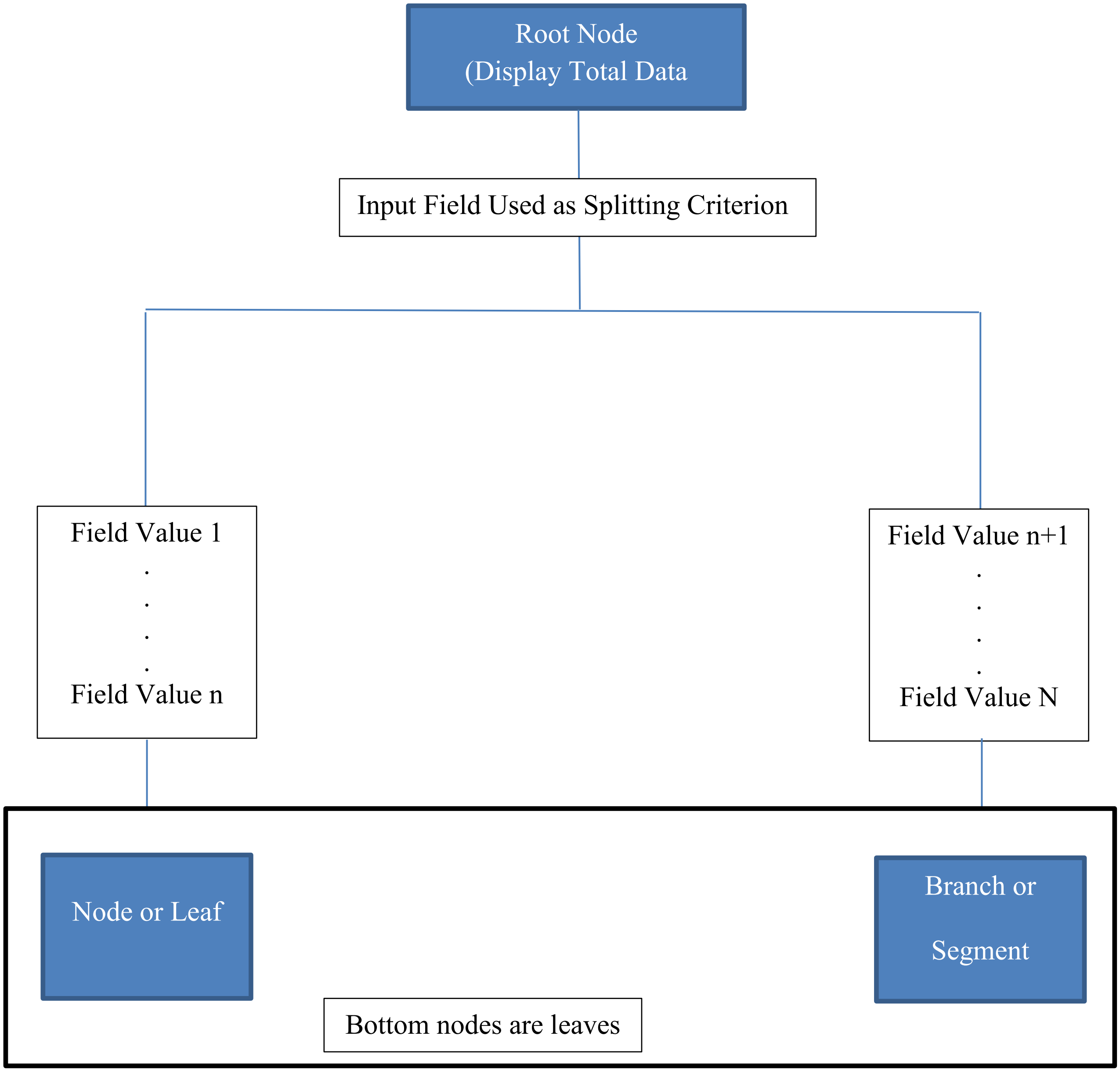


Figure 1.3: Illustration of the Decision Tree

In Chapter 5, the objective is to identify the stomach patients who could potentially benefit from surgery and those who could be very risk to take the surgery treatment, because the radiation or surgery may cause some side effect for the patients. We first perform the nonparametric and parametric survival analysis for comparing the two treatments' effect for male and female stomach cancer patients, respectively. Our result identified that the stomach patients who receive the combination of radiation and surgery have the most significant effect than the patients who receive only the radiation treatment with respect to the survival time. However, the decision tree model gives us the more powerful result. Based on the decision tree analysis, we found the more detailed treatment difference between different subgroups of the stomach cancer patients. For instance, a male stomach patient aged 70 to 76 years old with malignant tumor size between 40 and 58 millimeters, the combination of radiation and surgery shows the better effects on the survival time. Such important information could assist stomach cancer physicians to choose the suitable treatment for stomach cancer patients.

Chapter 2

Parametric Analysis of Malignant Stomach Tumor Size

2.1 Introduction

In medical research, the tumor refers to the cells' changing abnormality that is not necessarily meaning the body tumor. Some body parts of the cells will result in uncontrolled proliferation by the disease. Benign and malignant tumors are two types of tumors.

Benign tumors grow slowly and have smooth surface. They do not inbreak the normal tissue cells. A membrane usually envelops around the tumor body, so that it distinguishes between normal and abnormal cells. Normally, benign tumors will not lead to death and most of them can be completely removed.

Malignant tumors are cancerous and made up of cells that grow out of control. They break nearby tissues and spread to other parts of the body and distinguishes themselves from benign tumors. Sometimes cells move away from the primary cancer site to other organs and bones where they continue to grow and form another tumor at that site.

There is extensive research about the tumors of the stomach, especially cancer tumors. The research only provides descriptive information but doesn't clearly explain the changes in the disease rates, among others. This studies objective is exploring the causes for stomach cancer, among others. It is important to understand the natural history of the stomach cancer. In respect to this objective, we chose the malignant tumor sizes as the response variable and identified a statistical behavior of the tumor size. In the present study, we performed parametric analysis on the malignant tumor size of stomach cancer datasets obtained from the Surveillance, Epidemiology and End Results (SEER) database from 2004 up to 2013. We proceeded to identify the probability distribution functions that characterize the probabilistic behavior of malignant tumor size as a function of gender.

2.2 Data Description

A total of 11,462 records of stomach cancer patients from the year of 2004 to 2013 were obtained from SEER program. The SEER database is a premier source for various types of cancers in the United States. The information about cancer registries covers 28% of the population in the USA by collecting the data on all cancers that have been diagnosed. The SEER cancer statistics, which is a report of the most recent cancer incidence, mortality, survival, prevalence, and lifetime risk statistics, is published by the Surveillance research program of National Cancer Institute. The scope and purpose of this work are consistent with a report to the Senate Appropriations Committee (Breslow, 1988), which recommends that a broad profile of cancer can be presented to the American public on a routine basis.

Figure 2.1, given below, gives a characteristic network and presentation of the data that we will be working in the present study. This data is broken down to malignant and benign stomach cancer patients and as a function of gender and race.

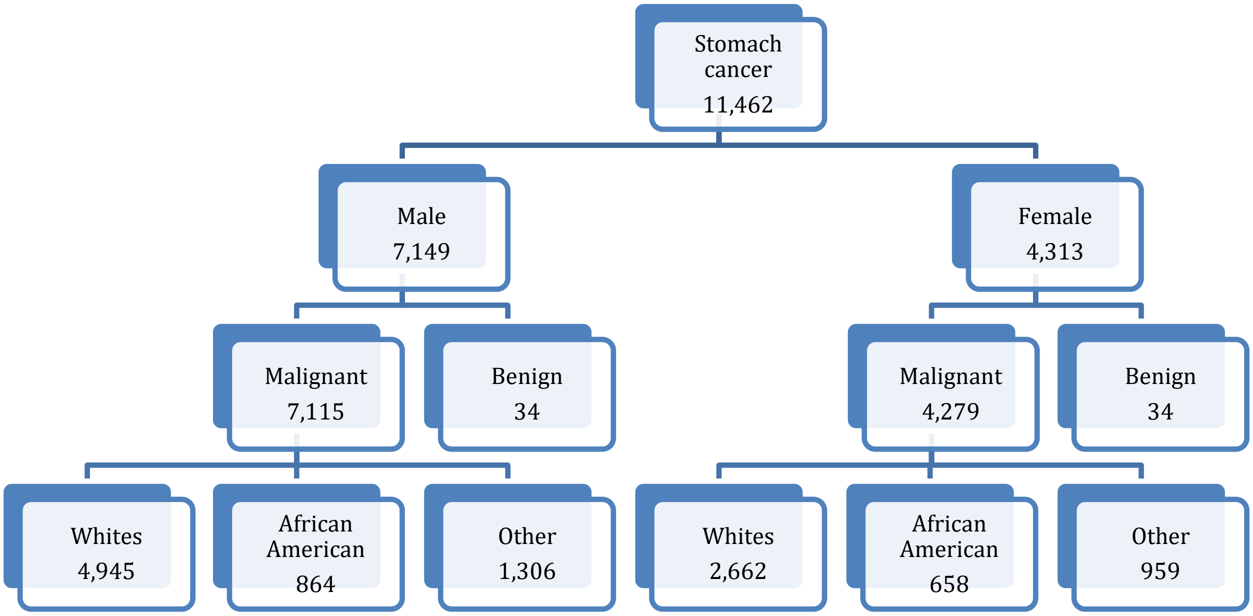


Figure 2.1: Schematic diagram of stomach cancer patients with malignant and benign tumor sizes for males and females.

In this study, we analyzed data on malignant primary stomach tumors diagnosed from 2004 to 2013 for the following registries: San Francisco-Oakland, Connecticut, Metropolitan Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah and Metropolitan Atlanta. The dataset includes 11,394 primary malignant stomach tumor records and 68 benign stomach tumor records. This analysis focuses only on the malignant tumors measured in millimeters.

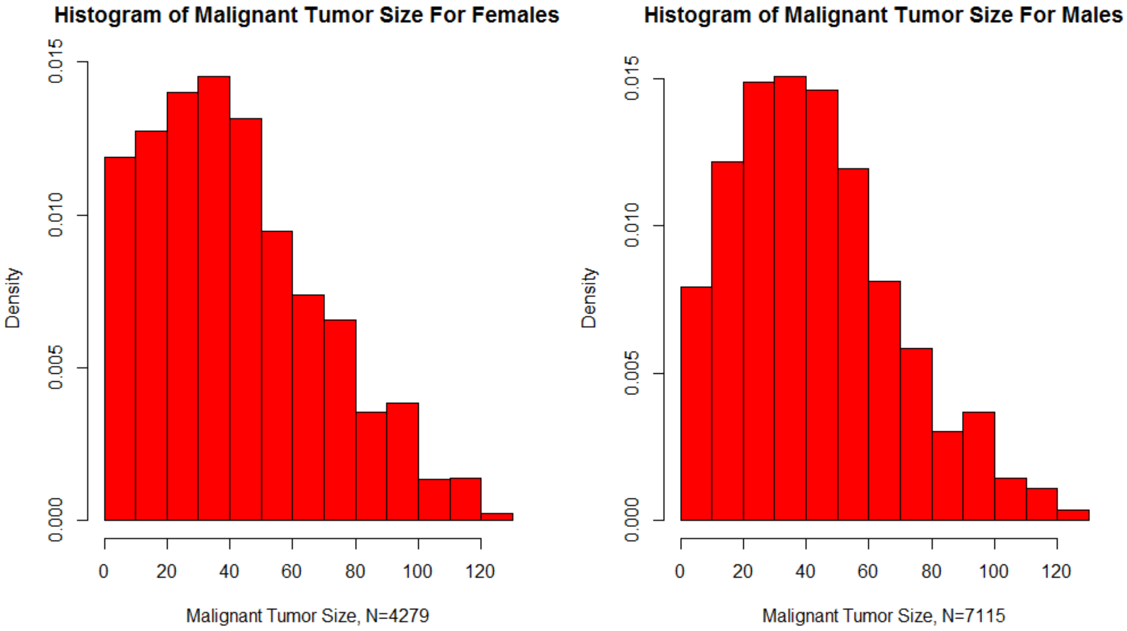
Using the information from the given data, we address the following questions:

1. What is the probability distribution function that characterizes the behavior of the malignant tumor sizes for each patient group?
2. Is there a significant difference in the mean cancerous tumor sizes between male and female patients?
3. If the answer to the second question is "yes", is there still a significant difference of the mean tumor sizes for male and female among different race groups (White, African American, Other)?

The answers to the above questions are important on how we proceed with inferential statistical analysis and modeling of the data.

2.3 Statistical Analysis of Gender and Race

Although Figure 2.2 below shows that the frequencies histogram plots of male and female are very similar, we still performed the parametric two-sample t-test and the nonparametric Kruskal-Wallis test to test for significant difference between the two genders. Based on the extremely small values, we rejected the null hypothesis at a level of significance of 0.0001 and concluded that there is a significant different between the mean malignant tumor sizes of male and female stomach cancer patients.


Histogram of Malignant Tumor Size For Females

Histogram of Malignant Tumor Size For Males

Figure 2.2: Frequencies Histogram plots of malignant tumor size for Female (left) and Male (right) patients.

Furthermore, we tested the mean of the malignant tumor size for males and females in different race groups (White, African American, Other), respectively. The extremely small values indicate that there is a significant difference among race groups at a significance level of 0.0001. In the next section, we introduce the best-fitted probability distribution functions for each different race group and by gender, respectively.

2.4 Probabilistic Behavior of Malignant Stomach Tumor Sizes

Parametric analysis is performed to determine the best fitted probability distribution function that characterizes the malignant tumor size behavior by gender and race.

The most important point of the parametric method is that we can use a small set of parameters to represent a large amount of information for the data. We can also get useful information from the estimated parameters. Therefore, we can use probabilistic models of malignant tumor sizes to explain the biological interpretation, among others. Furthermore, the useful information can be directly drawn from the estimated parameters for different subgroups, provided that the differences in the estimated parameters have a clear biological interpretation. Thus, one of the certain properties required for the probabilistic models of distributions of malignant tumor sizes is their biological interpretation. To better performing the analysis, we have partitioned the data set with regard to races of White, African American and the other races as well as the patient gender.

After about 50 different classical parametric probability distributions were used to fit the data, we obtained the best-fitted probability distributions. We applied three commonly used goodness-of-fit tests, i.e., Kolmogorov-Smirnov (Stephens, 1974) test, Anderson-Darling (T.W. Anderson, 1952) test and Chi-Square (H. Chenoff, 1954) test to determine the best probability distribution functions that characterize the malignant tumor sizes for male and female patients in different race groups. Finally, we identified that the best-fitted probability distribution functions that characterize the malignant tumor sizes are the Wakeby (Houghton, 1978) probability distribution for white female, African American female and other female race, the three-parameter Weibull (Cohen, 1965) probability distribution for white male and African American male, and the Dagum (Kleiber, 2008) probability distribution for other male race patients, respectively.

2.5 Parametric Analysis of Malignant Tumor Sizes for Female

2.5.1 Parametric Analysis of Malignant Tumor Sizes for White Female

We have found that the Wakeby probability distribution is the best-fitted probability distribution function that characterizes the malignant tumor size for white female patients with stomach cancer. The Wakeby probability distribution function is given by

$$f(x) = \frac{\left(1-F(x)\right)^{\delta+1}}{\alpha t+\gamma} + \xi, 0 < x < \infty \qquad (2.1)$$

where β, γ and δ are continuous shape parameters, ξ and α are location parameters, F is the cumulative probability distribution function given by $t = \left(1 - F(x)\right)^{\beta+\delta}$.

. The following conditions are imposed:

1. $\alpha \neq 0$ or $\gamma \neq 0$,
2. $\beta + \delta > 0$ or $\beta = \gamma = \delta = 0$,
3. If $\alpha = 0$, then $\beta = 0$,
4. If $\gamma = 0$, then $\delta = 0$,
5. $\gamma \geq 0$ and $\alpha + \gamma \geq 0$.

By using the maximum likelihood estimation, we found that the approximate maximum likelihood estimators of the parameters are,

$\hat{\alpha}$=58.697, $\hat{\beta}$=3.826, $\hat{\gamma}$=39.374, $\hat{\delta}$=-0.28324 and $\hat{\xi}$=-0.36842.

The identified Wakeby probability density function and quantile probability distribution function for white female patients are as follows:

$$f(x) = \frac{\left(1-F(x)\right)^{-0.28324+1}}{58.697t+39.374} - 0.36842, \tag{2.2}$$

where F is the cumulative probability distribution function and $t = \left(1 - F(x)\right)^{3.826-0.28324}$. The quantile function $x(F)$ is given by

$$x(F) = -0.36842 + \frac{58.697}{3.826}\left(1-(1-F)^{3.826}\right)$$
$$- \frac{39.374}{-0.28324}\left(1-(1-F)^{0.28314}\right). \tag{2.3}$$

We utilized the maximum likelihood estimates from Wakeby probability distribution to plot the estimated probability density curve of malignant tumor sizes for white female patients. Figure 2.3 illustrates that the identified distribution curve has a long right tail and most of the malignant tumor sizes are within the range of 0 to 120 millimeters. We plotted the estimated cumulative probability distribution function for white female patients as shown in Figure 2.4. The cumulative probability distribution function is explaining the characterize behavior of malignant tumor size. It assists us to find out the probability of the quantiles of the random variables, among others. For example, we can easily find that about 80% of white female patients have malignant tumor sizes less than or equal to 65 millimeters, etc.

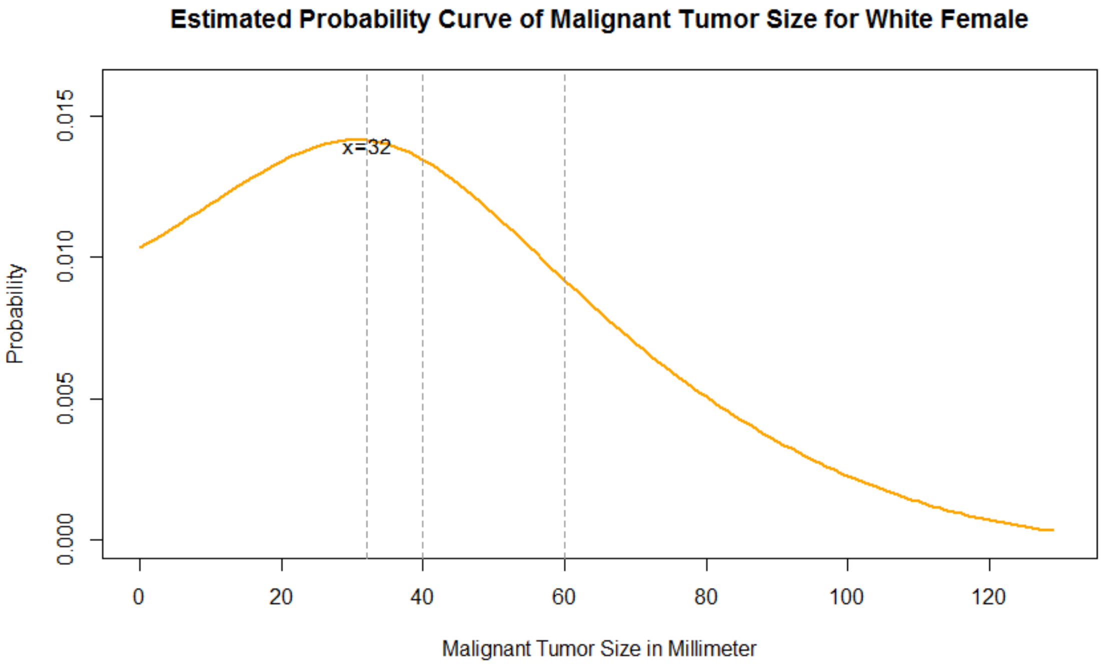


Figure 2.3: Fitted Wakeby Probability Distribution Function of Malignant Tumor Size for White Female

From Figure 2.3, we can see that the approximate probability of a white female patient has 32 millimeters malignant tumor size is 0.0142. And the approximate probability of a white female patient has the malignant tumor size between 40 millimeters and 60 millimeters is 0.235 that is $p(40 < x < 60) = 0.235$. And the probability of a white female patient has the malignant tumor size is greater than 40 millimeters is about 0.495.

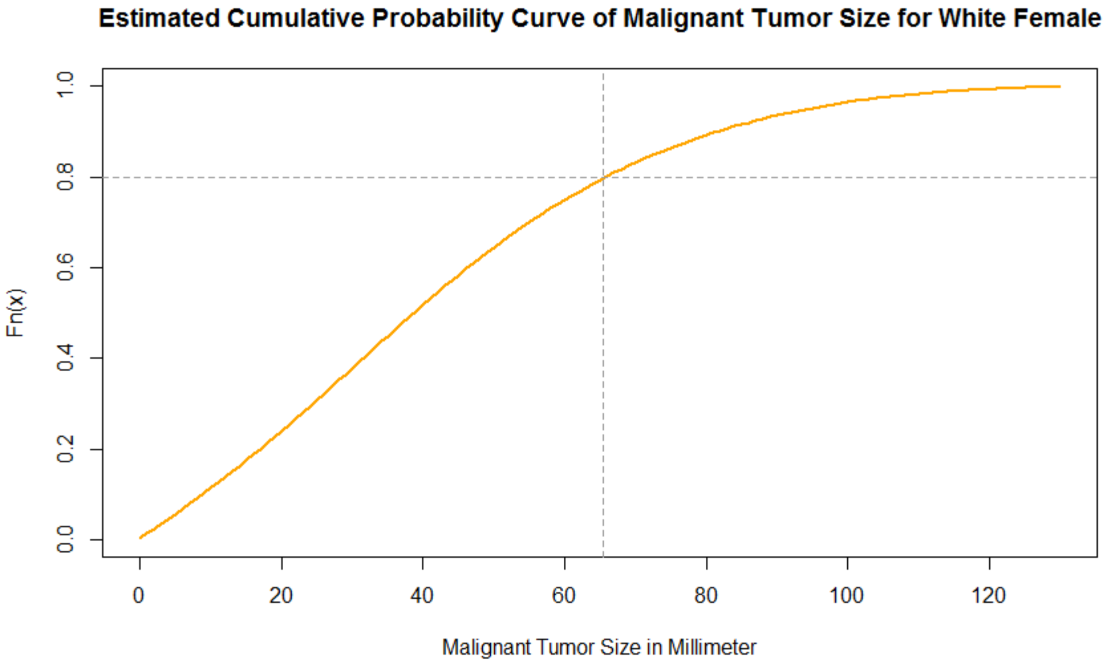


Figure 2.4: Fitted Cumulative Wakeby Distribution Function (CDF) of Malignant Tumor Size for White Female

The important statistical information using the Wakeby probability distribution are, the expected malignant tumor size,

$$E(x) = \frac{\hat{\alpha}}{1+\hat{\beta}} + \frac{\hat{\gamma}}{1-\hat{\delta}} + \hat{\mu}, when\ \delta < 1. \tag{2.4}$$

and the variance of the malignant tumor size,

$$V(x) = \frac{\hat{\alpha}^2}{(1+\hat{\beta})^2(1+2\hat{\beta})} - \frac{2\hat{\alpha}\hat{\gamma}}{(1+\hat{\beta})(1+\hat{\beta}-\hat{\delta})(-1+\hat{\delta})}$$

$$-\frac{\hat{\gamma}^2}{(-1+\hat{\delta})^2(-1+2\hat{\delta})}, when\ \hat{\delta} < 0.5. \tag{2.5}$$

and the 95% confidence interval of the true malignant tumor sizes is given by

$$P(a < true\ tumor\ size < b) \geq 0.95. \qquad (2.6)$$

We have obtained the expected mean of the malignant tumor size to be 42.48 millimeters for white female patients with a standard deviation of 27.64 millimeters. Moreover, we are at least 95% confident that to the true malignant tumor sizes for white female patients' to be between 2.041 millimeters and 105.09 millimeters with at least 95% confidence.

2.5.2 Parametric Analysis of Malignant Tumor Sizes for African American Female

After a very extensive search, we have identified that the Wakeby probability distribution is the best-fitted probability distribution function that characterizes the malignant tumor size for African American female patients with stomach cancer. The theoretical Wakeby probability distribution was introduced in section 2.4.1. The approximate maximum likelihood estimates of the five parameters are

$$\hat{\alpha}\text{=128.48, } \hat{\beta}\text{=8.3385, } \hat{\gamma}\text{=46.569, } \hat{\delta}\text{=-0.34268 and } \hat{\xi}\text{=-0.91885.}$$

Thus, the Wakeby probability density function that characterize the probabilistic behavior of the malignant tumor size of African American female patients is given by

$$f(x) = \frac{(1-F(x))^{-0.34268+1}}{128.48t+46.569} - 0.91885, < \mathrm{x} \qquad (2.7)$$

where F is the cumulative distribution function and $t = (1 - F(x))^{8.3385-0.34268}$. A graph of the estimated Wakeby probability distribution function for African American female patients with malignant stomach tumor size is given below by Figure 2.5.

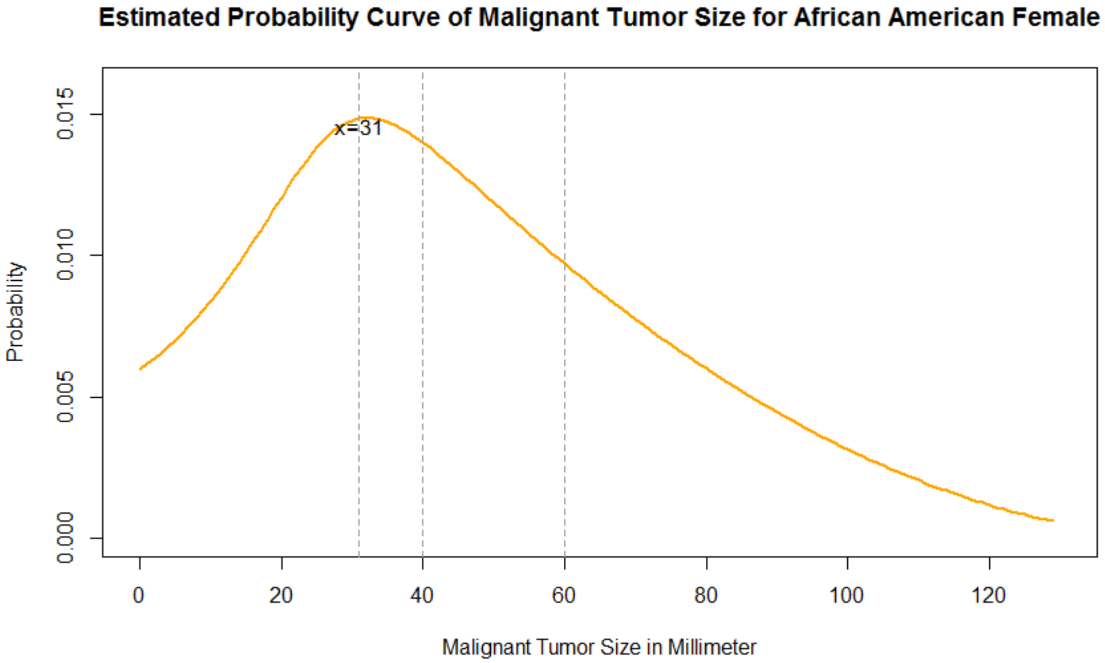


Figure 2.5: Fitted Wakeby Probability Distribution Function of Malignant Tumor Size for African American Female

The expected mean of the malignant tumor size was calculated to be 47.51 millimeters and the standard deviation to be 28.66 millimeters for African American female patients based on functions (2.4) and (2.5). This graph illustrates that the approximate probability of an African American female patient has 31 millimeters malignant tumor size is 0.0146. And the approximate probability of an African American female patient has the malignant tumor size between 40 millimeters and 60 millimeters is 0.243 that is $p(40 < x < 60) = 0.243$. And the probability of an African American female patient has the malignant tumor size is greater than 40 millimeters is about 0.558.

Figure 2.6 shows that about 80% of African American female patients have malignant tumor sizes less than or equal to 73 millimeters. Furthermore, we are at least 95% confidence to conclude that African American female patients' malignant tumor size fall in the interval (4.51 millimeters, 110.01 millimeters) by using the function (2.6).

Furthermore, the Wakeby cumulative probability distribution function is given by

$$x(F) = -0.91885 + \frac{128.48}{8.3385}(1 - (1 - F)^{8.3385}) - \frac{46.569}{-0.34268}(1 - (1 - F)^{0.34268}). \qquad (2.8)$$

Thus, a graphical form of F(x) is given below by Figure 2.6.

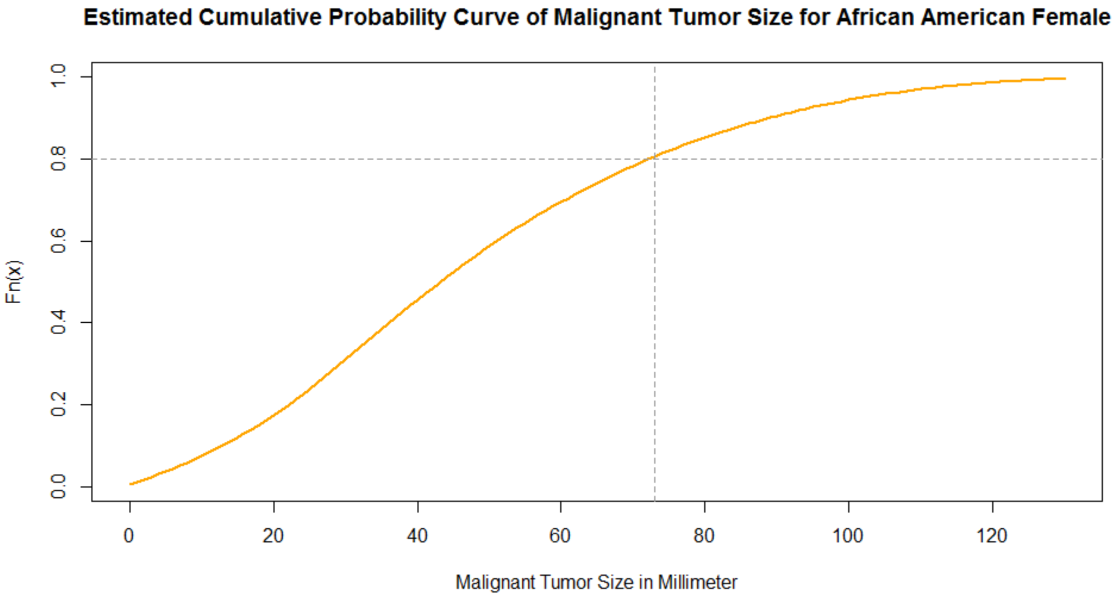


Figure 2.6: Fitted Cumulative Wakeby Distribution Function (CDF) of Malignant Tumor Size for African American Female

2.5.3 Parametric Analysis of Malignant Tumor Sizes for Other Female

Using the three popular goodness-of-fit tests that is, the Chi-square test, the Kolmogorov Smirnov test and the Anderson Darling test, we have identified that the Wakeby probability distribution also fits the malignant tumor size data for all other female race patients. The approximate estimates of the five parameters that are given by

$$\hat{\alpha}=108.16,\ \hat{\beta}=9.5265,\ \hat{\gamma}=48.728,\ \hat{\delta}=-0.38916 \text{ and } \hat{\xi}=-0.27848.$$

Thus, the identified Wakeby probability density function and quantile probability distribution function for other female race patients are as follows:

$$f(x) = \frac{\left(1-F(x)\right)^{-0.38916+1}}{108.16t+48.728} - 0.27848 \qquad (2.9)$$

where F(x) is a cumulative probability distribution function and $t = \left(1 - F(x)\right)^{9.5265-0.38916}$. The quantile probability distribution function is given by

$$x(F) = -0.27848 + \frac{108.16}{9.5265}(1-(1-F)^{9.5265})$$
$$-\frac{48.728}{-0.38916}(1-(1-F)^{0.38916}) \qquad (2.10)$$

The graphical form of (2.9) and (2.10) are given by Figures 2.7 and 2.8, respectively.

Figure 2.7 below illustrates that the approximate probability of an other race female patient has 27 millimeters malignant tumor size is 0.0152. And the approximate probability of an other race female patient has the malignant tumor size between 40 millimeters and 60 millimeters is 0.231 that is $p(40 < x < 60) = 0.231$. And the probability of an other race female patient has the malignant tumor size is greater than 40 millimeters is about 0.516.

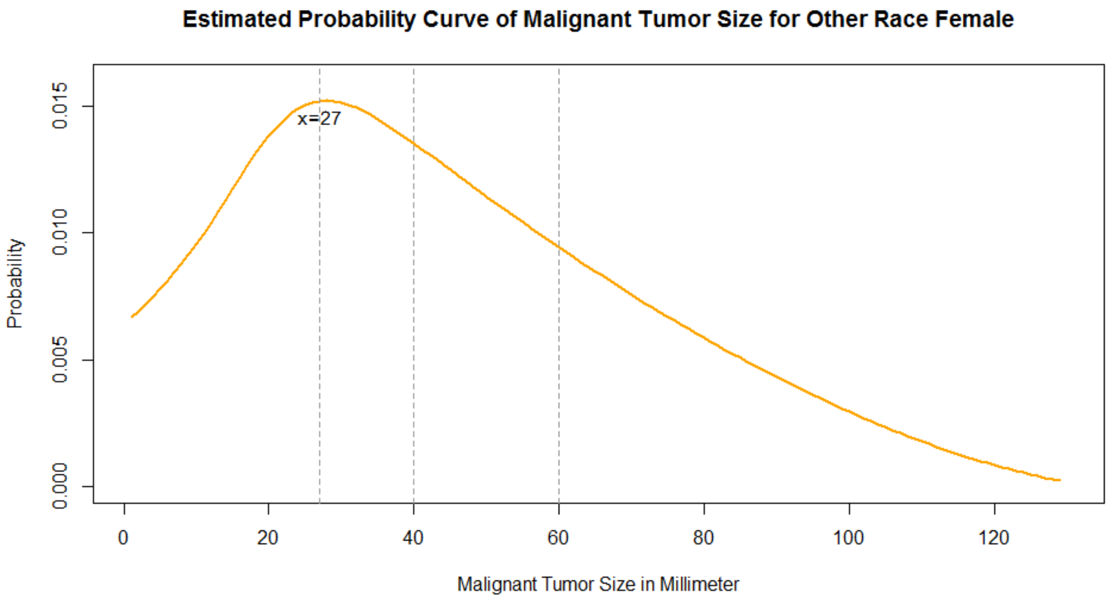


Figure 2.7: Fitted Wakeby Probability Distribution Function of Malignant Tumor Size for Other Female

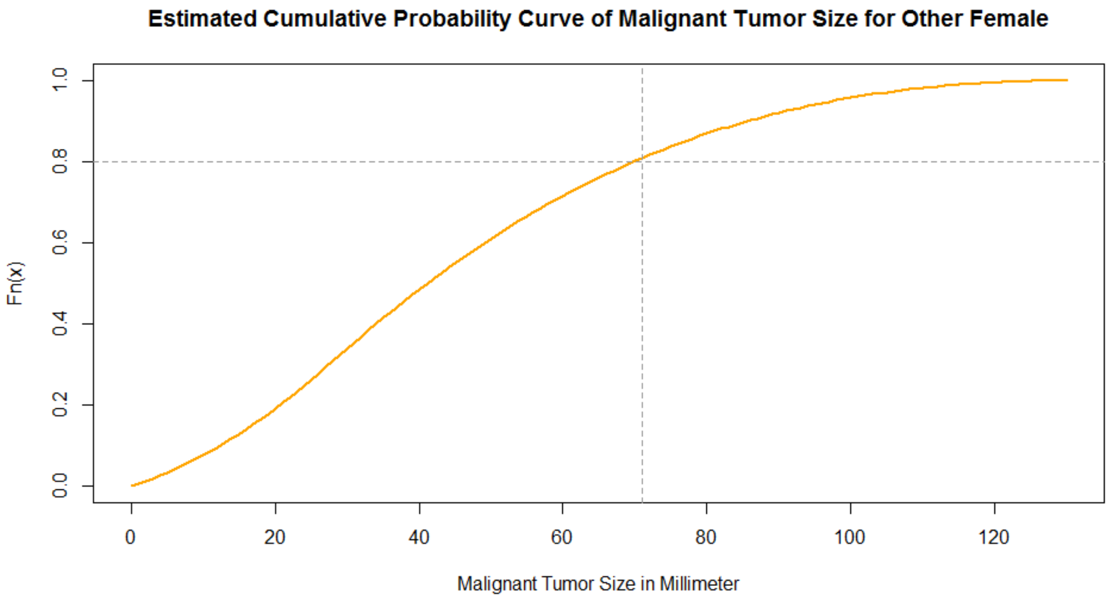


Figure 2.8: Fitted Cumulative Wakeby Distribution Function (CDF) of Malignant Tumor Size for Other Female

From Figure 2.8 above, we can estimate that about 80% of other female race patients have a malignant stomach tumor sizes less than or equal to 70 millimeters.

We also calculated the estimated expected mean of the malignant tumor size to be 45.56 millimeters and the standard deviation to be 27.96 millimeters for other female race patients. Furthermore, we are at least 95% confident that the true mean of the other female race patients' stomach malignant tumor size fall in the interval (5.27 millimeters, 110.85 millimeters) by using E.q (2.6).

2.5.4 Comparison of the Malignant Tumor Sizes for Female Patients

In this section, we are comparing the malignant tumor size for female patients by different race group. Table 2.1 below displays the best-fitted distributions that characterize the malignant tumor size for different race groups and the estimated malignant mean tumor size for female patients in each race group. We found the Wakeby probability distribution is the best fitted distribution that can characterize the behavior of the malignant tumor size for female patients in different race group. When we test the average malignant tumor size under different races (white, African American and other), we reject the null hypothesis that they are equal. The African American female patients have the largest expected malignant tumor size, which is 47.51mm. The white female stomach cancer patients have the smallest expected malignant tumor size, which is 42.48mm. Confidence intervals both provide the probabilistic behavior of the malignant tumors for each race group. This information will assist physicians to make appropriate decisions with a given degrees of assurance. Also, additional information such as confidence limits on the true size of the malignant tumor size will be useful to the physicians. For example, for a white female patient we are at least 95% confident her malignant tumor size is within the range from 2.04mm to 105.09mm and her expected malignant tumor size is 42.48mm.

Table 2.1: statistics of the malignant tumor sizes and confidence intervals for female patients

	PDF	Expected Mean	Standard Deviation	95% Confidence Intervals
White	Wakeby pdf	42.48	27.64	(2.041,105.09)
African American	Wakeby pdf	47.51	28.66	(4.51, 110.01)
Other Race	Wakeby pdf	45.56	27.96	(5.27, 110.85)

We plotted the estimated probability distribution curves of malignant tumor size of female patients for different race groups as shown in Figure 2.9 below. When the malignant tumor size is less than 20 millimeters, the white female patients have higher probability than the African American female and other female race patients do. When the malignant tumor size is between 20 millimeters and 40 millimeters, the other female race patients have higher probability. When the malignant tumor size is greater than 40 millimeters, the three race groups of patients have almost the same probability distributions. And the approximate expected true malignant tumor size for white female, African American and other female race patients are 42.48mm, 47.51mm and 45.56mm, respectively.

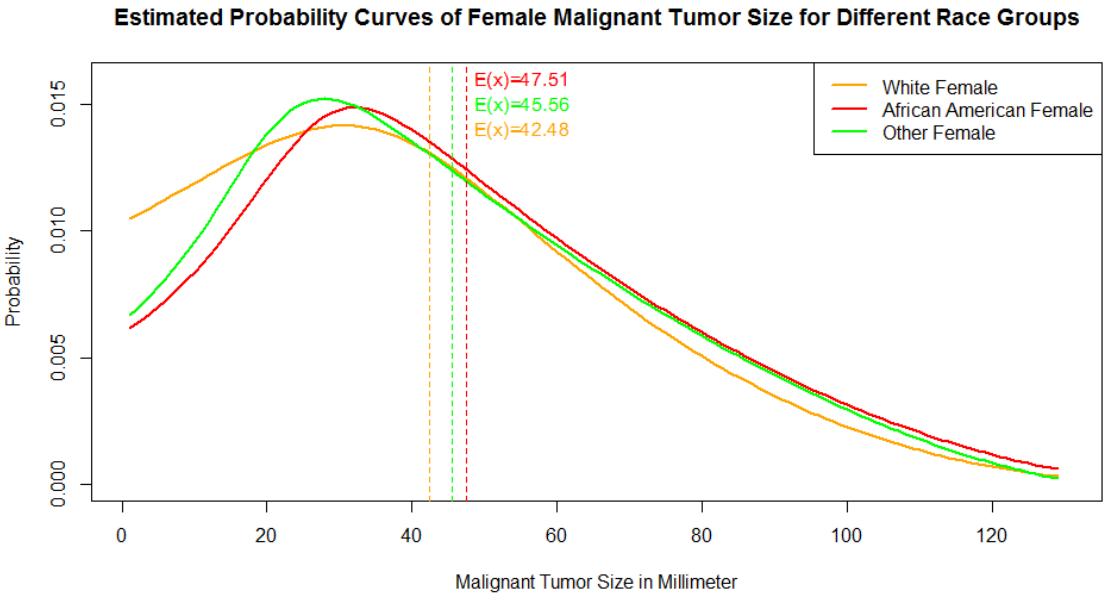


Figure 2.9: Fitted Distribution Curves of Malignant Tumor Size for Female Patients

Figure 2.10 below depicts the cumulative probability distribution curves of malignant tumor size for different female race groups. The orange (white female) curve is slightly higher than the other two lines. The green (other female race) line locates in the middle, which means that for a fixed value of malignant tumor size, the cumulative probability of white female patients is always greater than that of the other two race patients, and the cumulative probability of other female race patients is between that of white female and other female race. Moreover, the expected means of the malignant tumor size of female patients are 42.48 millimeters for white female, 47.51 millimeters for African American female and 45.56 millimeters for the other female race patients. The 95% confidence intervals are (2.041, 105.09) for white female, (4.51, 110.01) for African American female and (5.27, 110.85) for the other female race patients. These information is useful for doctor to convey to their patients.

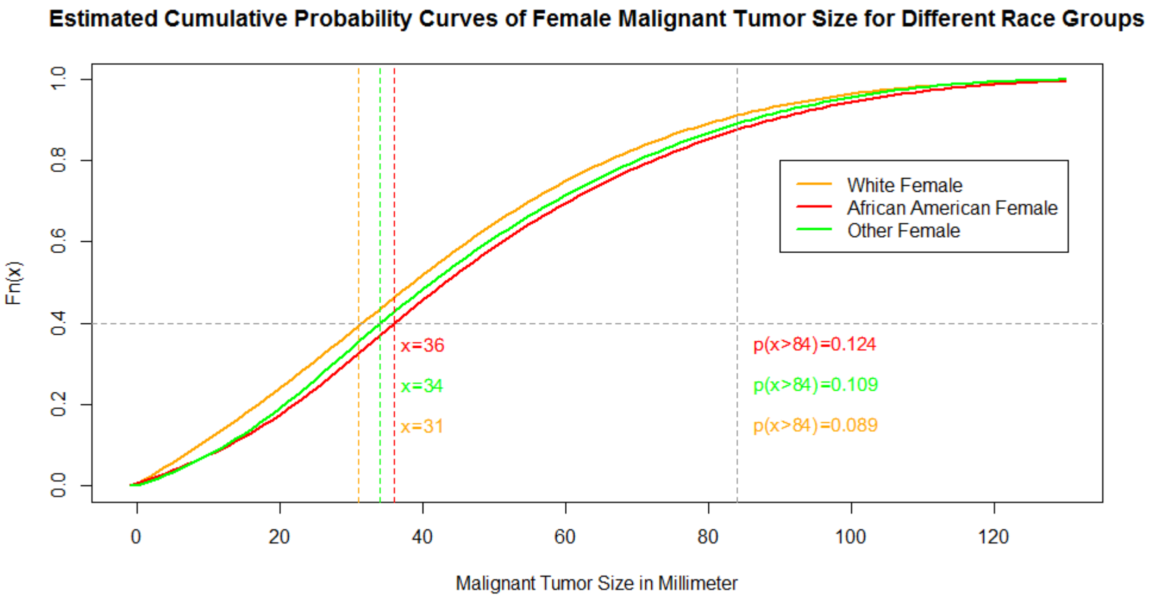


Figure 2.10: Fitted Cumulative Distribution Curves of Malignant Tumor Size for Female Patients

The above graph shows that the probability of the malignant tumor size is greater than 84mm are 0.089, 0.124 and 0.109 for white female, African American female and other female race patients, respectively. Furthermore, there are 40% of the stomach cancer patients have the tumor size less than 31mm, 36mm and 34mm for white female, African American female and other female race patients, respectively.

2.6 Parametric Analysis of Malignant Tumor Sizes for Male

2.6.1 Parametric Analysis of Malignant Tumor Sizes for White Male

The Weibull probability distribution is one of the most commonly used probability distributions in health science studies. The Weibull probability distribution has the best fit of the malignant tumor sizes data for white male patients. The 3-parameter Weibull probability distribution and cumulative probability distribution functions are given by

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right), 0 < x \qquad (2.11)$$

And

$$F(x) = 1 - exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right), 0 < x \qquad (2.12)$$

where α, β, γ are the shape, scale and location parameters. When $\gamma = 0$, results in the commonly used two-parameter Weibull probability distribution.

Based on the Chi-square, Komogorov Smirnov and Anderson Darling tests, we identified the Weibull probability distribution be the best fitted probability distribution for white male patients. The approximate maximum likelihood estimates of the parameters that drive the Weibull probability distribution function are $\hat{\alpha}$=1.7966, $\hat{\beta}$=51.362, $\hat{\gamma}$=-1.1904. Thus, the probability density function (pdf) and the corresponding cumulative probability distribution function (CDF) for the three-parameter Weibull probability distribution are respectively given by:

$$f(x) = \frac{1.7966}{51.362}\left(\frac{x+1.1904}{51.362}\right)^{1.7966-1} exp\left(-\left(\frac{x+1.1904}{51.362}\right)^{1.7966}\right) \qquad (2.13)$$

And

$$F(x) = 1 - exp\left(-\left(\frac{x+1.1904}{51.362}\right)^{1.7966}\right). \qquad (2.14)$$

The expected mean and variance of the malignant tumor size for white male patients are as follows:

$$E(x) = \hat{\gamma} + \hat{\beta}\Gamma\left(\frac{1}{\hat{\alpha}} + 1\right) \qquad (2.15)$$

$$= 44.48mm$$

and

$$V(x) = \hat{\beta}^{2}\left(\Gamma \quad \left(\frac{2}{\hat{\alpha}} + 1\right) \quad -\Gamma \quad \left(\frac{1}{\hat{\alpha}} + 1\right)^{2}\right) \qquad (2.16)$$

$$= 26.30$$

The graphs of $f(x)$ and $F(x)$ are given by Figures 2.11 and 2.12, respectively. Figure 2.11 below shows that the probability of a white male patient has 20mm and 40mm tumor sizes are 0.0137 and 0.01514, respectively. And the probability of a white male patients has the tumor size between 20mm and 40mm is 0.304 that is

$$p(20mm < x < 40mm) = 0.304.$$

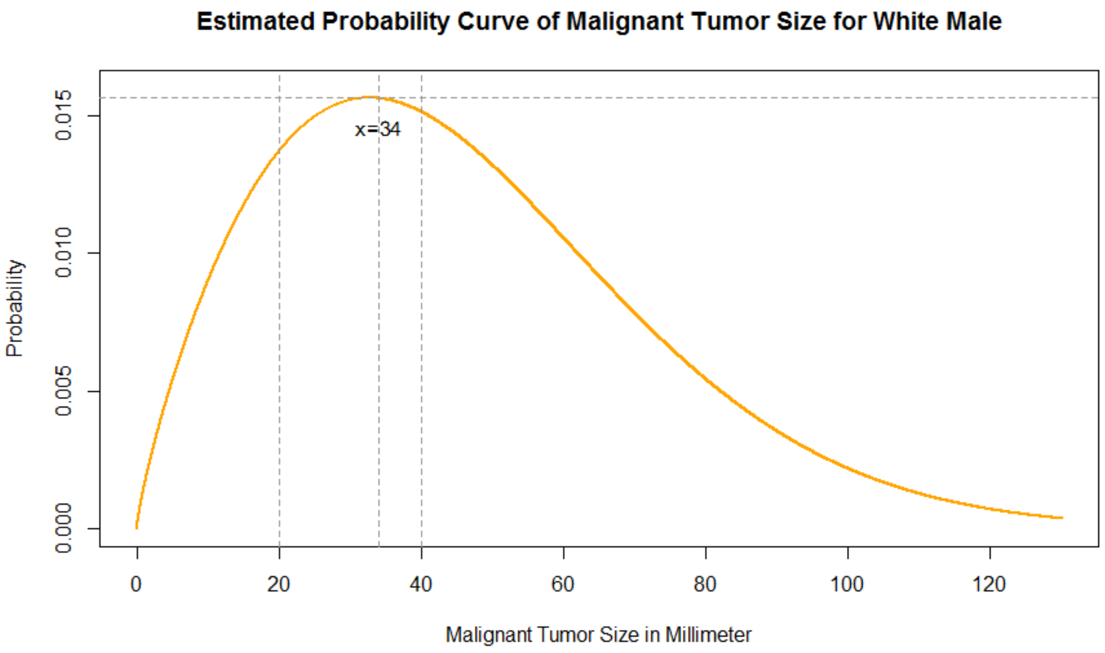


Figure 2.11: Fitted Weibull Probability Distribution Function of Malignant Tumor Size for White Male

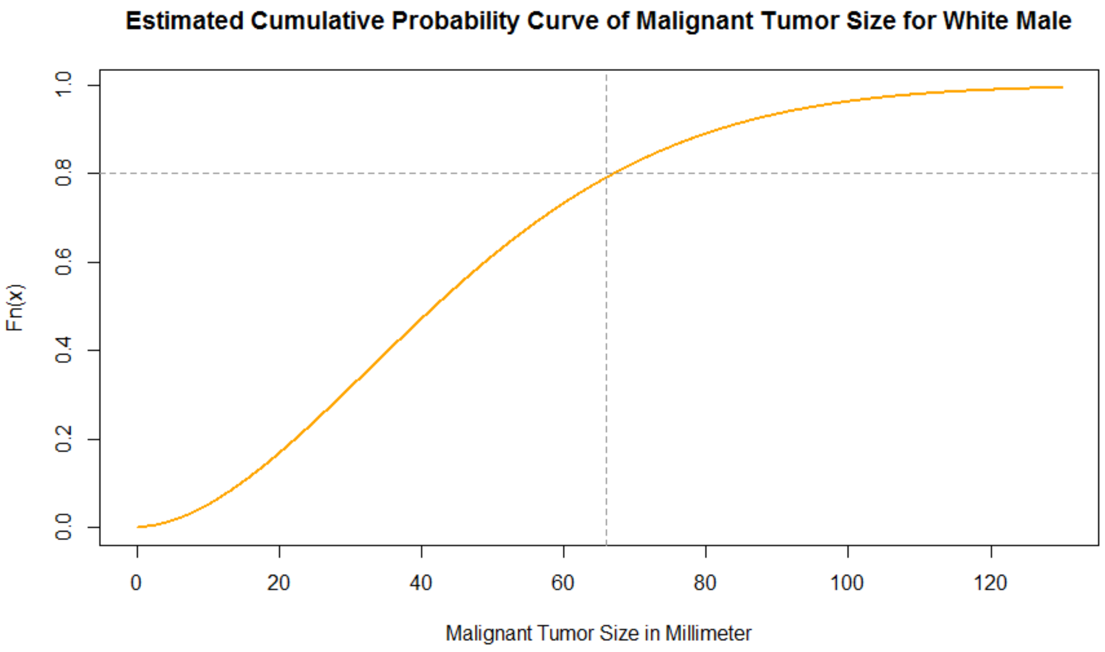


Figure 2.12: Fitted Cumulative Weibull Distribution Function (CDF) of Malignant Tumor Size for White Male

Thus, we calculated the expected mean of the malignant tumor size is 44.48 millimeters and the standard deviation for white male patients is 26.30 millimeters. We also obtained the confidence interval based on E.q (2.6), that is, we are at least 95% confident that white male patients' the true malignant tumor size is in the range from 3.17 millimeters to 104.66 millimeters, that is

$$p(3.17mm < true\ tumor\ size < 104.66) \geq 0.95$$

From Figure 2.12 above. It illustrates that around 80% of the white male patients have the malignant tumor size less than or equal to 66 millimeters. That is

$$p(X \leq 66mm) \approx 0.80.$$

2.6.2 Parametric Analysis of Malignant Tumor Sizes for African American Male

The results of the goodness-of-fit tests that is Chi-square test, Kolmogorov Smirnov and Anderson Darling test allow us to identify the best-fitted probability distribution function for African American male patients with stomach malignant tumor is the three-parameter Weibull probability distribution. We obtained the approximate maximum likelihood estimates of the three parameters are $\hat{\alpha}$=1.9248, $\hat{\beta}$=57.193, $\hat{\gamma}$=-1.2722. Thus, the fitted probability density function (pdf) and the corresponding cumulative distribution function (CDF) for African American Male patients are respectively given by:

$$f(x) = \frac{1.9248}{57.193}\left(\frac{x+1.2722}{57.193}\right)^{1.9248-1} exp\left(-\left(\frac{x+1.2722}{57.193}\right)^{1.9248}\right), 0 < x \qquad (2.17)$$

and

$$F(x) = 1 - exp\left(-\left(\frac{x+1.2722}{57.193}\right)^{1.9248}\right), 0 < x \qquad (2.18)$$

The graphs of the identified three-parameter Weibull probability distribution function (2.17) and cumulative probability distribution function (2.18) are given by Figures 2.13 and 2.14, respectively.

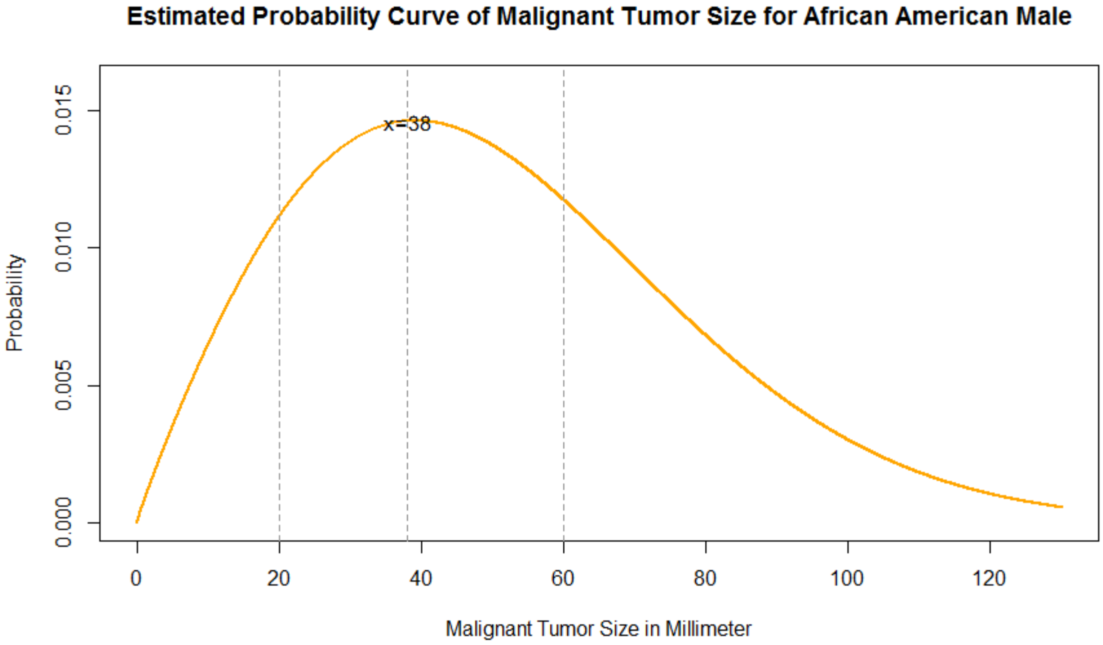


Figure 2.13: Fitted Weibull Probability Distribution Function of Malignant Tumor Size for African American Male

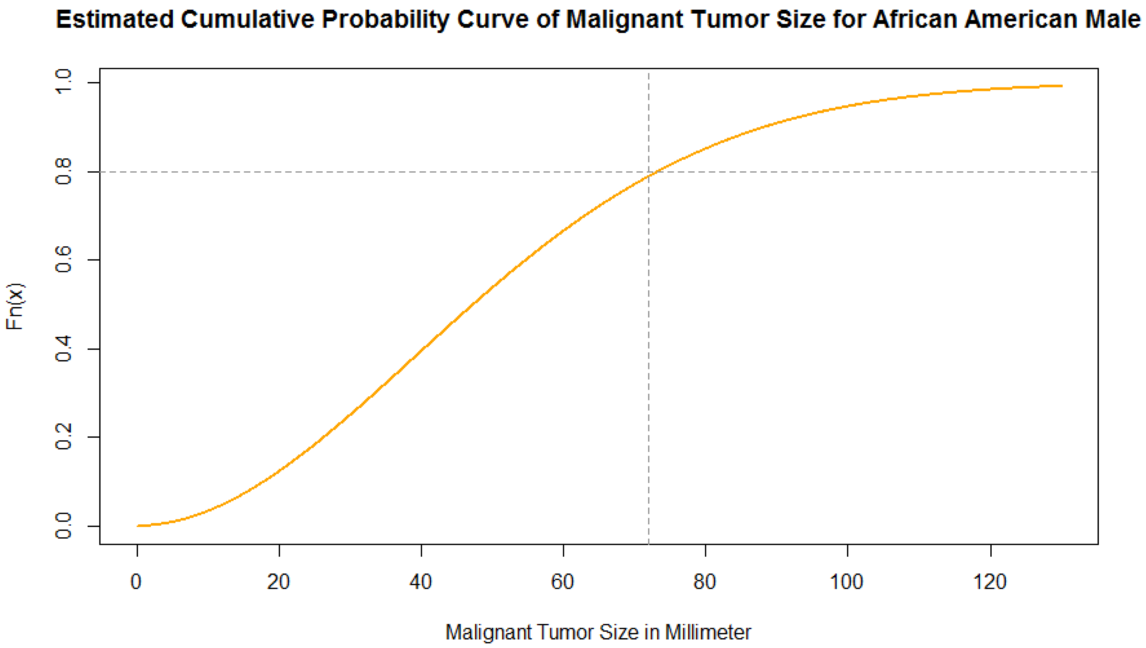


Figure 2.14: Fitted Cumulative Weibull Distribution Function (CDF) of Malignant Tumor Size for African American Male

From Figure 2.14 above, we can obtain that around 80% of African American male patients have malignant tumor sizes less than or equal to 72 millimeters. Also, $p(20mm < x < 60mm) = 0.542$.

We also calculated the estimated expected mean of the malignant tumor size is 49.46 millimeters and the standard deviation for African American male patients is 27.45 millimeters. Furthermore, we are at least 95% confident that the true mean of the African American male patients' malignant tumor size is in the range from 7.2 millimeters to 111.41 millimeters. That is,

$$p(7.2mm < true\ tumor\ size < 111.41mm) \geq 0.95.$$

2.6.3 Parametric Analysis of Malignant Tumor Sizes for Other Male

Similarly, the results of goodness-of-fit tests, that is, Kolmogorov-Smirnov test (Stephens, 1974), Anderson-Darling test (T.W. Anderson, 1952) and Chi-Square test (H. Chenoff, 1954), we have identified that the best-fitted probability distribution function that characterizes the malignant tumor sizes for the other male race patients is the three parameter Dagum probability distribution.

Dagum probability distribution is a continuous probability distribution. It is named after Camilo Dagum, who proposed it in a series of papers in the 1970's. The Dagum probability distribution function and cumulative probability distribution function are shown by equation 2.19 and 2.20, respectively.

$$f(x; \alpha, k, \beta) = \frac{\alpha k \left(\frac{x}{\beta}\right)^{\alpha k-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{k+1}} \tag{2.19}$$

and

$$F(x) = \left(1 + \left(\frac{x}{\beta}\right)^{-\alpha}\right)^{-k} \tag{2.20}$$

Where $k, \alpha, \beta > 0$ and $\gamma < x < \infty$. α and k are the shape parameters, β is the scale parameter and γ is the location parameter. When γ=0, the four-parameter Dagum probability distribution reduce to the three-parameter Dagum probability distribution.

The approximate maximum likelihood estimates of the parameters that derive the probability distribution function are given by $\hat{\alpha}$=5.363, $\hat{\beta}$=71.69, $\hat{k}$=0.25867. Thus, the Dagum probability distribution function and its cumulative probability distribution are given below:

$$\mathrm{f(x)} = \frac{5.363*0.25867\left(\frac{x}{71.69}\right)^{5.363*0.25867-1}}{71.69\left(1+\left(\frac{x}{71.69}\right)^{5.363}\right)^{0.25867+1}}, 0 < x \qquad (2.23)$$

$$\mathrm{F(x)} = \left(1 + \left(\frac{\mathrm{x}}{71.69}\right)^{-5.363}\right)^{-0.25867}, 0 < x \qquad (2.24)$$

The expected mean and variance of the malignant tumor size for other male race patients are as follows:

$$E(X) = -\frac{\frac{\beta}{\alpha}\Gamma\left(-\frac{1}{\alpha}\right)\Gamma\left(\frac{1}{\alpha}+k\right)}{\Gamma(k)}, (\alpha > 1) \qquad (2.21)$$

$$= 46.88mm$$

And

$$V(X) = -\frac{\beta^2}{\alpha^2}\left(\quad 2\alpha\frac{\Gamma\left(-\frac{2}{\alpha}\right)\Gamma\left(\frac{2}{\alpha}+k\right)}{\Gamma(k)} + \quad \left(\frac{\Gamma\left(-\frac{1}{\alpha}\right)\Gamma\left(\frac{1}{\alpha}+k\right)}{\Gamma(k)}\right)^2\right), (\alpha > 2) \qquad (2.2)$$

$$= 789.04$$

The graphical form of the Dagum probability distribution function $f(x)$ and its cumulative distribution function $F(x)$ are given by Figures 2.15 and 2.16, respectively.

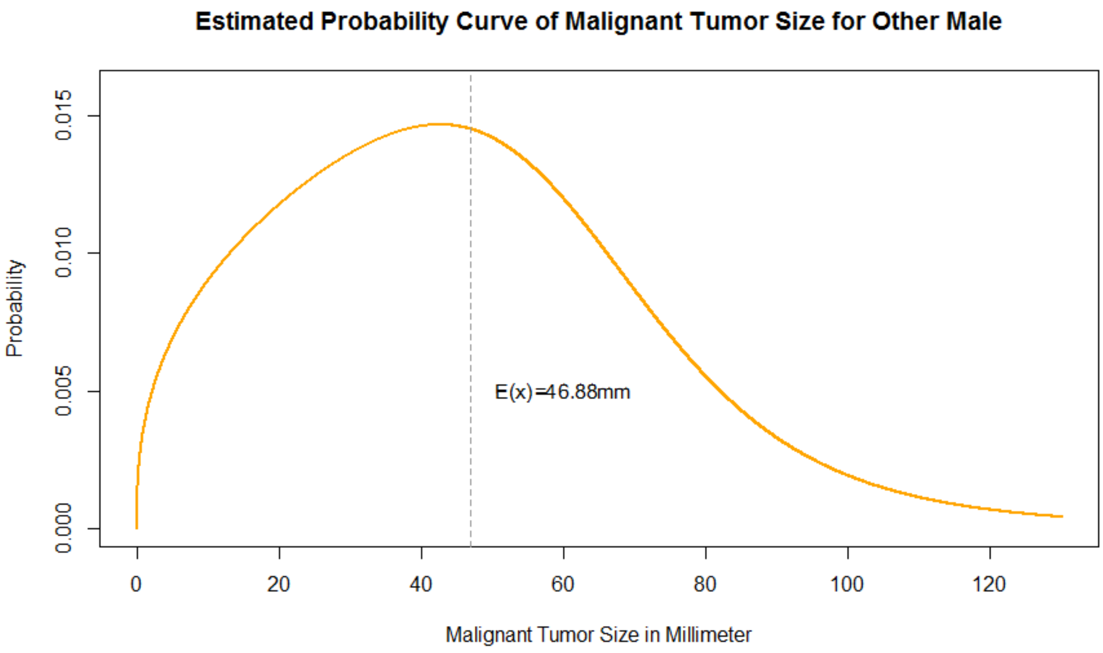


Figure 2.15: Fitted Dagum Probability Distribution Function of Malignant Tumor Size for Other Male

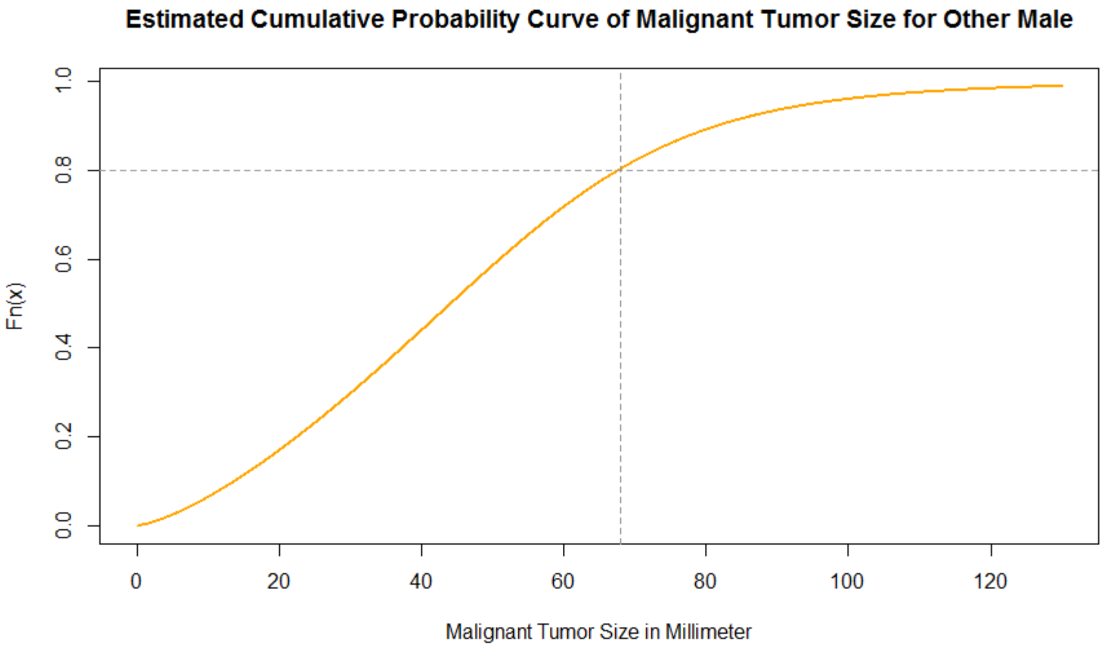


Figure 2.16: Fitted Cumulative Dagum Distribution Function (CDF) of Malignant Tumor Size for Other Male

From Figure 2.16 above. We can see that about 80% of the other male race patients have the stomach malignant tumor size less than or equal to 68 millimeters. Also, $p(x \geq E(x) = 46.88) = 0.459$, the probability that the malignant tumor size will be larger than the expected tumor size.

We also calculated the expected mean of the stomach malignant tumor size is 46.88 millimeters and the variance for other male race patients is 789.04 millimeters. We can use equation (2.6) to obtain confidence limits as the true size of the malignant tumor. That is,

$$p(5.02mm \leq true\ size \leq 109.56mm) \geq 0.95.$$

2.6.4 Comparison of the Malignant Tumor Sizes for Male Patients

In this section, we compared the malignant tumor size for male patients by different race group. Table 2.2 below displays the best-fitted distributions that characterize the malignant tumor size for different race groups and the estimated malignant mean tumor size for male patients in each race group. We found the Weibull probability distributions are the best fitted distributions that can characterize the behavior of the malignant tumor size for white and African American male patients, while the Dagum probability distribution for other male race patients. When we compare the average malignant tumor size under different races (white, African American and other), we reject the null hypothesis that they are all the same. The African American male patients have the largest expected malignant tumor size, which is 49.46mm, while the white male stomach cancer patients have the smallest expected malignant tumor size, which is 44.48mm. We obtained confidence intervals for the true malignant tumor size for all the cases. The confidence range is 101.49mm, 104.21 and 104.51mm for white male, African American male and other male, respectively. Such information is useful for the physicians.

Table2.2: statistics of the malignant tumor sizes and confidence intervals for male patients

	PDF	Expected Mean	Standard Deviation	95% Confidence Intervals
White	Weibull pdf	44.48	26.30	(3.17,104.66)
African American	Weibull pdf	49.46	27.45	(7.20, 111.41)
Other Race	Dagum pdf	46.88	28.09	(5.02, 109.56)

We plotted the estimated probability distribution curve of the three-parameter Weibull probability distribution for white male patients and African American male patients and the probability distribution curve of Dagum probability distribution for the other male race patients as shown by Figure 2.17 below. We can conclude that when the malignant tumor size is below 15 millimeters, the probabilities of white male and other male races are almost the same and they both are greater than that of African American male. When the malignant tumor size is between 15 millimeters and 45 millimeters, the probabilities of African American male and other male race are almost the same, and they have probability less than that of white male patients. When the malignant tumor size is greater than 45 millimeters, white male patients have lower probabilities and the other two groups cross each other.

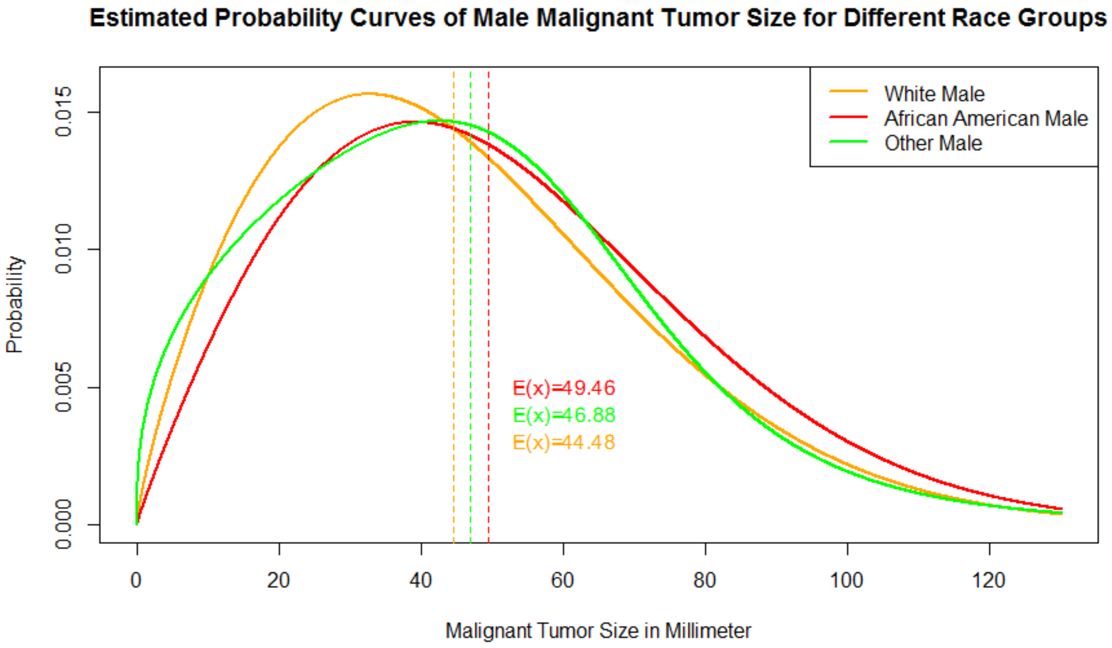


Figure 2.17: Fitted Probability Distribution Curves of Malignant Tumor Size for Male Patients

We plotted the cumulative probability distribution curves of malignant tumor size for the three different race groups for male patients as shown in Figure 2.18 below. When we fixed malignant tumor size, the white male patients always have higher cumulative probability than the other two groups. African American male patients have the smallest cumulative probability of malignant tumor size. The probability of the malignant tumor size is greater than 84mm are 0.089, 0.123 and 0.088 for white male, African American male and other male race patients, respectively. Furthermore, there are 40% of the stomach cancer patients have the tumor size less than 35mm, 40mm and 37mm for white male, African American male and other male race patients, respectively.

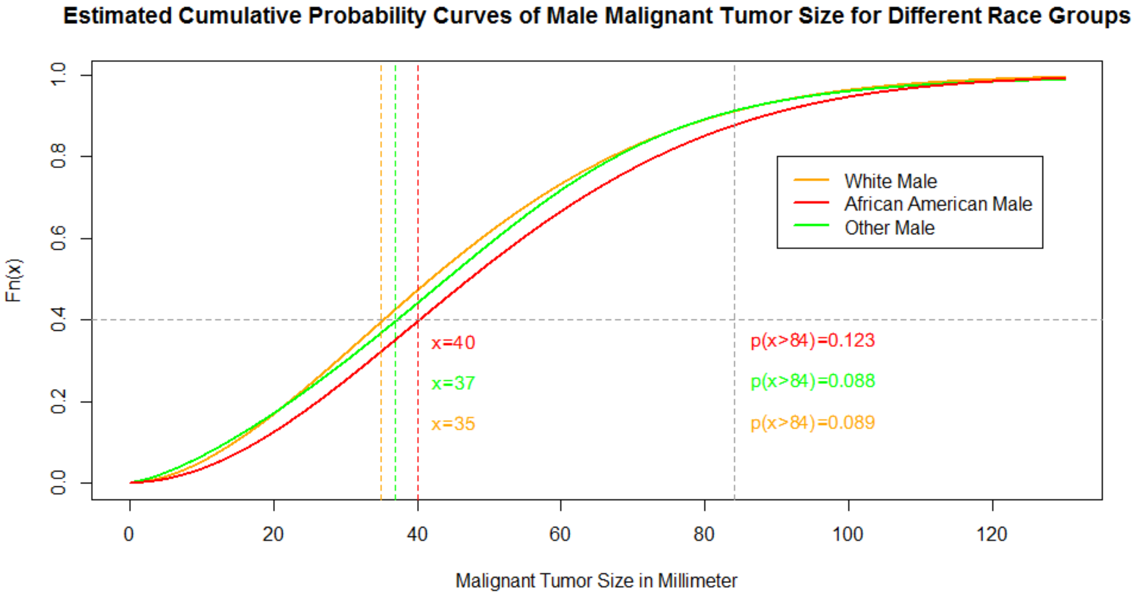


Figure 2.18: Fitted Cumulative Probability Distribution Curves of Malignant Tumor Size for Male Patients

The previous Table 2.2 demonstrates that the expected mean malignant tumor sizes are 44.48 millimeters for white male patients, 49.46 millimeters for African American male patients and 46.88 millimeters for the other male race patients. The corresponding 95% confidence intervals are (3.17, 104.66) for white male, (7.2, 111.41) for African American male and (5.02, 109.56) for other male race patients.

2.7 Relationship between Malignant/Non-Malignant Tumor and Tumor Size

In order to investigate the relationship between the malignancy of the tumor and the tumor size, we need to find out the chance that a stomach tumor becomes malignant. To better achieve this objective, we conducted a plot to exhibit the probability of a malignant tumor given a function of the tumor size. The details of the relationship are addressed in Figure 2.19 below.

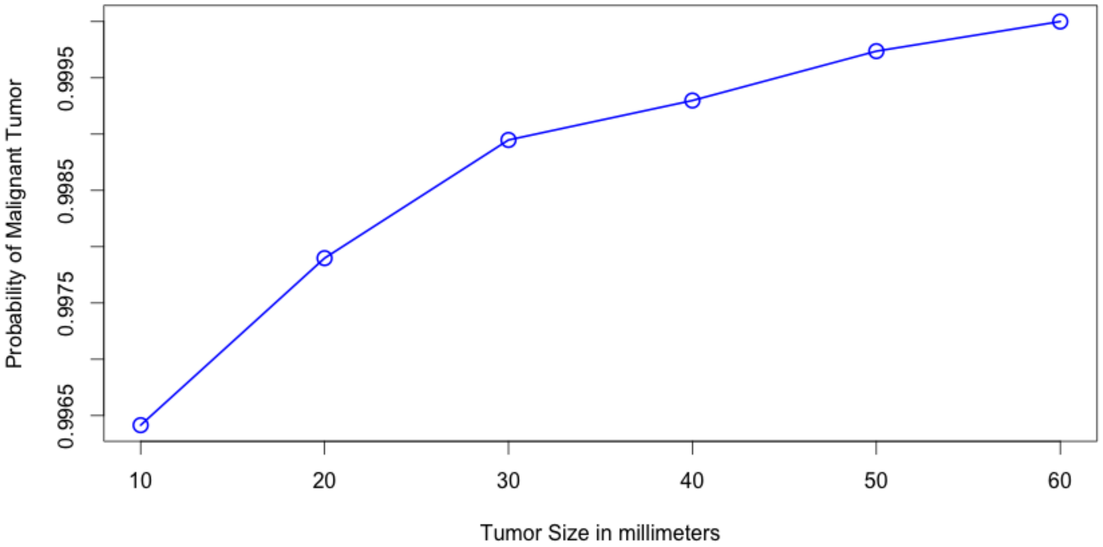


Figure 2.19: Probability of Malignant Tumor as a Function of Tumor Size

Figure 2.19 shows an increasing trend between the size of a tumor and the probability of the tumor to be malignant. When the tumor size is 10 millimeters, it has a 99.6% chance that the tumor is malignant. Furthermore, we could conclude that it has nearly a 99.99% chance that the tumor is malignant when the patient has a tumor size of 60 millimeters or bigger. The plot illustrates that a patient has an extremely high chance to have a malignant tumor although the size of the tumor is small.

2.8 Contributions

Table2.3: statistics of the malignant tumor sizes and confidence intervals for stomach cancer patients

	PDF		Expected Value		Standard Deviation		95% Confidence Intervals	
	F	M	F	M	F	M	F	M
White	Wakeby	Weibull	42.48	44.48	27.64	26.30	(2.041,105.09)	(3.17,104.66)
African American	Wakeby	Weibull	47.51	49.46	28.66	27.45	(4.51, 110.01)	(7.20,111.41)
Other Race	Wakeby	Dagum	45.56	46.88	27.96	28.09	(5.27, 110.85)	(5.02,109.56)

We have developed parametric analysis by defining the probability distributions of malignant tumor size for different race groups in the United States from 2004 to 2013. Table 2.3 above summarize the results of the parametric analysis from which we can obtain the following useful information.

1. We have demonstrated that the mean of cancerous tumor size is different for gender and race groups.
2. The best fitted probability distribution function for white female patients is Wakeby probability distribution with mean 42.48 mm and standard deviation 27.64 mm.
3. The best fitted probability distribution function for African American female patients is Wakeby probability distribution with mean 47.51 mm and standard deviation 28.66 mm.

4. The best fitted probability distribution function for other female race patients is Wakeby probability distribution with mean 45.56 mm and standard deviation 27.96 mm.
5. The best fitted probability distribution function for white male patients is Weibull probability distribution with mean 44.48 mm and standard deviation 26.30 mm.
6. The best fitted probability distribution function for African American male patients is the three parameter Weibull probability distribution with mean 49.46 mm and standard deviation 27.45 mm.
7. The best fitted probability distribution function for other race male patients is Dagum probability distribution with mean 46.88 mm and standard deviation 28.09 mm.
8. The fitted probability distribution functions are essential to develop statistical inference on the malignant tumor size.
9. The graphical figures of the probability density functions and the cumulative distribution curves could provide assistances for physicians understand the probabilistic behavior of the tumor size.

Chapter 3

Statistical Quantile Regression Model for Malignant Stomach Tumor

3.1 Introduction

The average malignant stomach tumor size is often insufficient to explain the probabilistic behavior of malignant stomach tumor size. Standard least square estimates only shows an average effect of the covariates on the average malignant stomach tumor sizes. Thus, it is desirable to investigate the best-fitted statistical model that will describe the effects of malignant stomach tumor sizes by using the covariates of race, gender, age and their interaction terms. As demonstrated in the parametric analysis (Chapter 2) of malignant stomach tumor size study, we found the more appropriate probability distributions to describe the probabilistic behavior of the malignant stomach tumor sizes are highly skewed and follow the Wakeby probability distribution for White female, African American female and other female race patients; the three-parameter Weibull probability distribution fits best for White male and African American male patients; and the Dagum probability distribution for all other male race patients. Besides the fact that regression curves can provide the summary of the average of the probability distribution, regression curves with different percentages are able to depict a more complete picture of the malignant tumor size. This chapter describes the effects of race, gender and age on the sizes of malignant stomach tumor by developing a statistical quantile regression model (R. Koenker, 1978).

The proposed statistical quantile regression model comprehensively describes the effects of the predictors on the response variable by modeling the relationship between a set of risk variables (i.e., age, gender, race and their interaction terms) and the specific percentiles of the response variable (i.e., malignant tumor size). For example, the 50th percentile of malignant stomach tumor size on patients' age, gender and race specifies the changes in the median malignant tumor size as a function of patients' race, gender, age and their interaction terms. The effect of the specific predictors (race, gender, age and their interaction terms) on the median malignant tumor sizes can be compared with its effect on the other quantiles of malignant tumor size. Statistical median regression is more robust to outliers than least squares regression, and it is semi parametric because it avoids assumptions for the parametric distribution of the error process. Statistical quantile regression models are widely used in many fields such as environmental sciences, econometrics, survival analysis, among others.

3.2 Data Description

In the formulation of the quantile statistical regression model for the stomach cancer data was obtained from Surveillance, Epidemiology, and End Results database (SEER). Figure 3.1 below shows a schematic diagram of the actual data that we used along with the appropriate covariates. We considered the races of white, African American and other; the gender of female and male and the age of the stomach cancer patients as risk factors, with the quantiles of the malignant stomach tumor sizes as the response variable. The malignant tumor size is measured in millimeters, and the total number of patients with malignant tumor size is 11,394. For the patients with the stomach malignant tumor, there are 7,115 males and 4,279 female patients. Moreover, there are 7,607 white patients, 1,522 African American patients along with 2,265 other race patients.

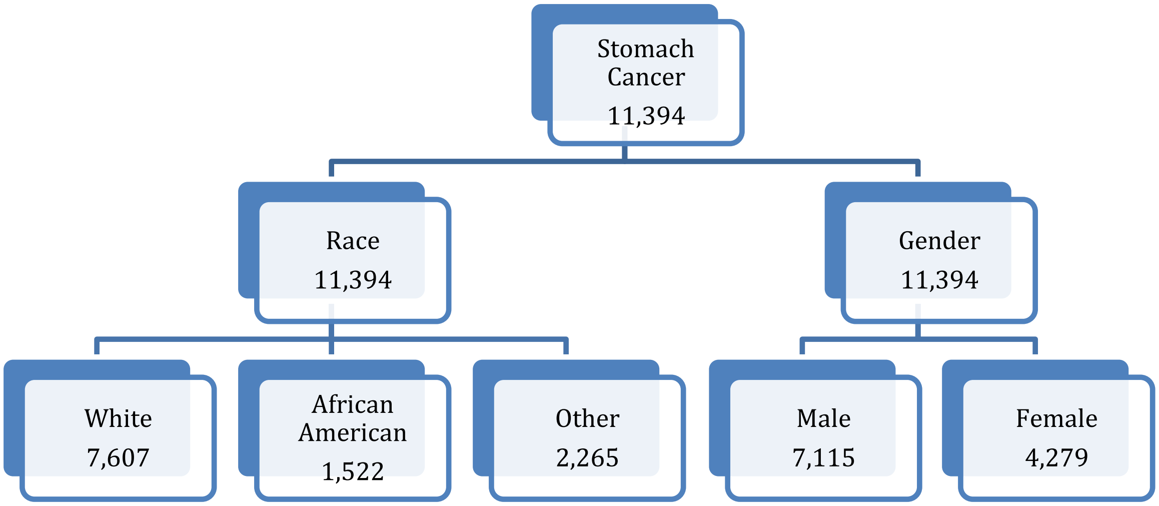


Figure 3.1: Schematic diagram of stomach cancer patients

3.3 Stomach Tumor Sizes and Quantile Regression Model

3.3.1 Quantiles and Optimization

The quantile regression model is a conditional quantile given by the predictors. For a random variable Y with a cumulative probability distribution function given by

$$F(y) = Prob\ (Y \leq y). \tag{3.1}$$

The τ^{th} quantile of Y is denoted as the inverse function (Koenker 2005), that is,

$$F^{-1}(\tau) = Q(\tau) = \inf\{y: F(y) \geq \tau\}, \tag{3.2}$$

where $0 < \tau < 1$. More specifically, the median regression quantile is given by

$$Q(0.5)\ = \inf\{y: F(y) \geq 0.5\}. \tag{3.3}$$

Fox and Rubin (1964) used the piece-wise linear loss function to estimate the quantile estimators. They also used the ordered sample observations to verify the τ^{th} quantile.

For a random sample $\{y_1, y_2, \dots, y_n\}$ of Y, the sample median can be computed by minimizing the following equation which is the sum of the absolute deviations, that is,

$$\text{Sample median}=min_{\varepsilon \in R} \sum_{i=1}^{n} |y_i - \varepsilon|. \tag{3.4}$$

Similarly, the τ^{th} sample quantile $\varepsilon(\tau) = Q(\tau)$ can be computed as a solution of the optimization problem, that is,

$$Q(\tau) \ = min_{\varepsilon \in R} \sum_{i=1}^{n} \rho_\tau(y_i - \varepsilon), \tag{3.5}$$

where $\rho_\tau(u) = u\big(\tau - I(u < 0)\big), 0 < \tau < 1$, and $I(.)$ is the indicator function.

We proceed to minimize the following sum of squared residuals,

$$\text{Sample mean} = \hat{\mu} = argmin_{v \in R} \sum_{i=1}^{n} (y_i - \mu)^2. \tag{3.6}$$

By solving

$$\hat{\beta} = \ argmin_{\beta \in R^p} \sum_{i=1}^{n} (y_i - x'_i\beta)^2, \tag{3.7}$$

we can extend (3.6) to conditional expectation $E(Y = X = x) = x'\beta$.

Similarly, we can solve the following equation

$$\hat{\beta} = \ argmin_{\beta \in R^p} \sum_{i=1}^{n} \rho_\tau(y_i - x'_i\beta), \tag{3.8}$$

to estimate the linear conditional quantile function $Q(\tau|X = x) = x^T\beta(\tau)$,

where $\hat{\beta}(\tau)$ is denoted an estimate as the τ^{th} regression quantile for any quantile $\tau \in (0,1)$. There is a special case when $\tau = 0.5$ which minimizes the sum of absolute residuals.

3.3.2 Regression Quantile Estimates

Let $X = (x_1, x_2, \dots, x_n)$ be a random variable and $Y = (y_1, y_2, \dots, y_n)$ be the corresponding n observed responses. The statistical quantile regression model is then defined by

$$y = X^T \beta_\tau + \varepsilon_\tau, \tag{3.9}$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ denotes an unknown vector of parameters and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$ represents a vector of unknown errors terms. The τ^{th} regression quantile can be computed as a solution of

$$min_{\beta \in R} \left[\sum_{y_i \geq x_i^T \beta} \tau \left| y_i - x_i^T \beta \right| + \sum_{y_i < x_i^T \beta} (1 - \tau) \left| y_i - x_i^T \beta \right| \right], \tag{3.10}$$

and we can obtain the median regression by solving the following function,

$$min_{\beta \in R} \left[\sum_{y_i \geq x_i^T \beta} \frac{\left| y_i - x_i^T \beta \right|}{2} + \sum_{y_i < x_i^T \beta} \frac{\left| y_i - x_i^T \beta \right|}{2} \right], \tag{3.11}$$

which is a special case of (3.10).

In the literature, statisticians are able to estimate the coefficients of the median regression as a linear programming problem by solving special forms of the simplex algorithm (Barrodale and Roberts 1973). The simplex algorithm is widely used in many statistical scientific areas. Although the computation loops exponentially increase as the sample size increases, scientists commonly apply the simplex algorithm to the data with sample size less than 10,000. Alternative approaches have been developed for large data sets. The interior point approach (Karmarkar 1984) is widely used to solve median regression problems in which the relevant interior of a constraint set is approximated by an ellipsoid. The interior point approach has been proven to be better than the simplex algorithm as well as dealing with large data sets. Excluding the simplex algorithm and interior point method, there are several heuristic estimations that have been developed to explore the median quantile regression solutions. The most powerful method is the finite smoothing algorithm (Madsen and Nelsen 1993). The Newton-Ralphon algorithm utilizes the finite number of loops to find the parameter coefficients since its smoothing algorithm approximates the objective regression function with a smoothing function.

To measure the effects of the patients' age, gender, race and their interaction terms on the malignant stomach tumor sizes, we have developed statistical quantile regression models for different quantiles of the distributions of malignant tumor sizes. In this section, we give a brief introduction for the proposed model and apply the proposed model to the SEER stomach cancer data set.

The standard least square regression models only compute the average mean effect of independent variables on the tumor size. Standard least squares assumption of normally distributed errors does not hold for stomach cancer data base because the malignant tumor sizes follow Dagum, Weibull and Wakeby probability distributions. Thus, to focus only on the average malignant tumor size, we use the standard least square regression models that probably hide some important elements of the underlying relationship. Statistical quantile regression has been used most commonly to model the skewed data in the literature.

In medicine, reference charts provide a collection of the useful quantiles. Comparing with a simple reference range, quantile curves have merits when the measurement strongly depends on a covariate such as patient's age (Cole and Green 1992, Royston and Altman 1994). In survival analysis, a given covariate may have different effects at different quantile levels. Such effects can be solved by using statistical quantile functions on survival time (Koenker and Geling, 2001). Statistical quantile regression model can also be applied in the field of economics (Buchinsky 1998, Machado and Mata 2005). Pokhrel (2013) discussed the effect of brain tumor sizes by quantile regression model. Unfortunately, there is no such statistics analysis for stomach cancer until the present study.

There are several merits that lead us to choose the statistical quantile regression model rather than the ordinary least squares estimation:

a. The ordinary least squares are inefficient when the errors are highly non-normally distributed. Since the stomach malignant tumor sizes follow the Dagum, Wakeby and Weibull probability distributions, the standard least squares assumption of normally distributed errors fails to hold.

b. Statistical quantile regression model is invariant to monotonic transformations and the coefficients $\hat{\beta}_\tau$ are invariant to outliers of dependent variable (Buchinsky, 1994).

c. Quantile regression models can comprehensively describe the conditional probability distribution of dependent variables (Alex Coad and Rekha Rao, 2008).

d. The probability distribution of malignant stomach tumor sizes is skewed as revealed in our previous study. Statistical quantile regression model is robust to the highly skewed probability distributions.

Thus, we proposed the statistical quantile regression methods to model the relationship between a set of risk factors (i.e., race, gender, age and their interaction terms) and particular quantiles of the response variable (i.e., malignant stomach tumor size) by specifying the changes in the quantiles of malignant stomach tumor size. For example, a median quantile regression of malignant tumor size diagnosed on the malignant stomach tumor patients specifies the changes in the median tumor size as a function of the predictors. Similarly, the 25th quantile regression of malignant tumor size diagnosed on patients specifies the changes in the 25th quantile of tumor size as a function of the covariates. The effect of patients' age on the 75th quantile malignant stomach tumor sizes of white patients can be compared with that of African American patients. Such information exhibits a more comprehensive picture of the race effect on age and the race effect on the upper or lower side of the distribution of malignant tumor sizes. Furthermore, in linear regression models, the regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding the other predictors fixed. By modeling the estimates with a particular quantile of the response variable for one unit of change in the risk factor, quantile regression allows us to find the percentiles of malignant stomach tumor size that are affected more by the specific characteristics of patients.

We proceeded to calculate the coefficient estimates at 19 different quantiles of the conditional probability distribution through the following statistical quantile regression model:

$$Q_\tau = \alpha + \beta_1 Race + \beta_2 Sex + \beta_3 Age + \beta_4 Age * Sex + \beta_5 Age * Race + \beta_6 Sex * Race + \varepsilon, \qquad (3.12)$$

where α is the model intercept, β_i's are the coefficients and ε is the error term.

The ordinary least squares estimates require restrictive assumptions of the error terms, which are identically distributed from normal distribution. By avoiding that, we proceed to use the statistical quantile regression model on the stomach cancer data which we obtained from SEER database.

The τ^{th} sample quantile can be obtained by solving the following minimization problem

$$\widehat{Q_\tau} = argmin_{q \in R}[(1-\tau)\sum_{y_i<q}(q-y_i) + \tau \sum_{y_i \geq q}(y_i - q)], \qquad (3.13)$$

Where q is an initial guess for τth quantile of the sample data (Q_τ).

Based on the point estimate $\widehat{Q_\tau}$, we would also want to introduce a 95% confidence interval for the desired population quantile. In the literature, the interpolated order statistic approach suggested by Hettmansperger and Sheather (1986) and Nyblom (1992). The distribution free method has the advantage of robustness against studentization method because it does not need to assume the normality of the error terms (Kenneth, Zhou, 1996). The interval should fulfill the following condition:

$$p\left(x_{(d)} \leq x_p \leq x_{(e)}\right) \geq 1 - \alpha.$$

Then we need to try out all possible (d, e) combinations where d, e satisfy the following conditions:

$$p\left(x_{(d)} \leq x_p\right) \geq 1 - \frac{\alpha}{2},$$

and

$$p\left(x_p \leq x_{(e)}\right) \geq 1 - \frac{\alpha}{2}.$$

If several combinations achieve our criterion, we need to calculate the length of each intervals and then take the minimal length.

Figures 3.2-3.5 display the results of our statistical quantile regression for the stomach cancer data set. In each plot, we utilized 5th to 95th quantiles for the regression coefficients, which indicate the effect on the malignant stomach tumor size with one unit change in that variable while other covariates remain the same. The shade regions are bands of the point wise with at least 95% confidence intervals. For example, Figure 3.2 below represents the estimated conditional quantile function of the malignant stomach tumor size of “other” female race. If the other race female patient has 40 millimeters malignant tumor size then she will be fall in the 50th quantile. On the other side, if we know the other race female patient falls in the 75th quantile, then we have at least 95% confident to conclude that she may have the malignant tumor size between 56 millimeters and 61 millimeters.

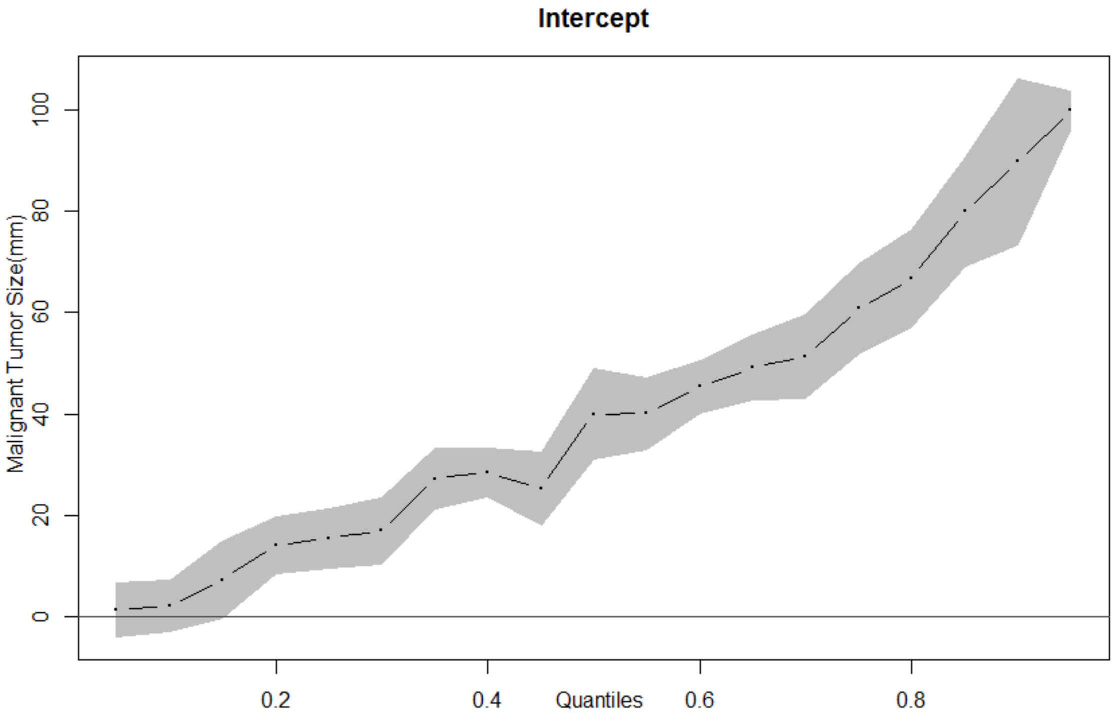


Figure 3.2: The estimated intercept parameter by quantile regression for malignant tumor size. The shaded area depicts at least 95% point-wise confidence band

Figure 3.3 below displays the estimated difference of malignant stomach tumor size between white patients and other race stomach cancer patients. It shows the negative effect of malignant stomach tumor size when comparing with other race patients. For instance, for 60th quantile, we have at least 95% confident that the white patients will have the tumor size less than that of other race patients, the difference will between 0 and 13 millimeters.

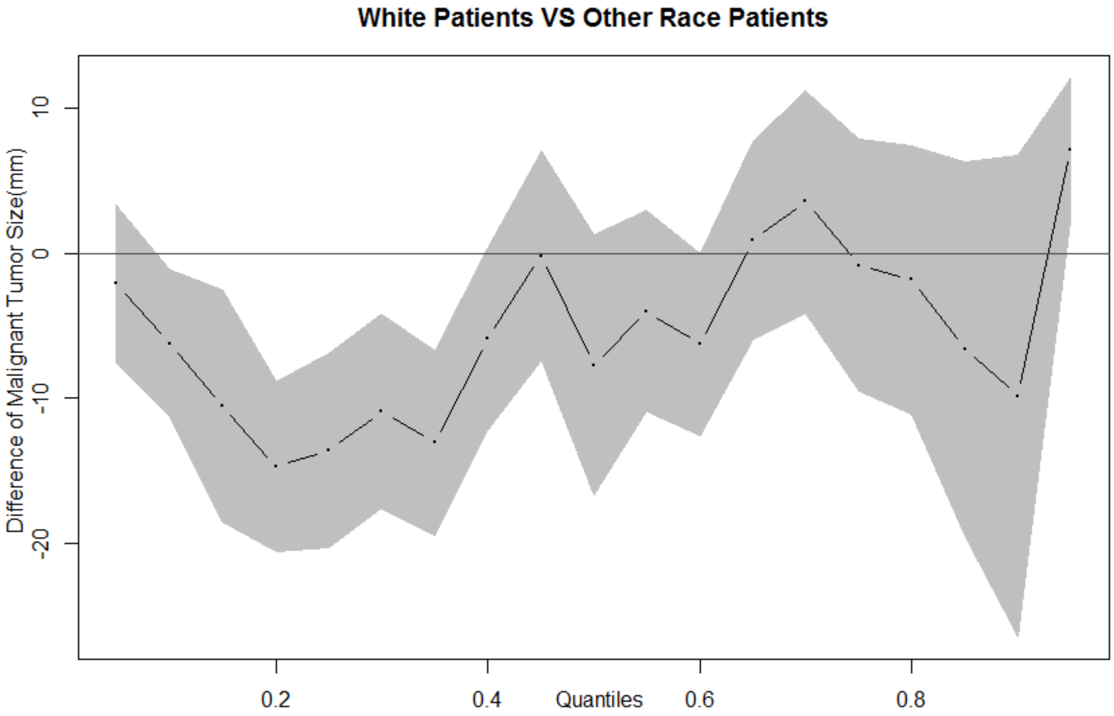


Figure 3.3: The estimated difference of malignant tumor size between white and other race patients. The shaded area depicts at least 95% point-wise confidence band

Figure 3.4 below displays the estimated difference of malignant tumor size between African American patients and other race patients by using our proposed quantile regression model. It shows the negative effect of malignant stomach tumor size for low quantiles when comparing with other race patients, as well as a positive effect for upper quantiles. For instance, for 20th quantile, the African American patients will have the smaller tumor size than that of other race patients and we have at least 95% confident that the difference will between 1.5 and 16 millimeters.

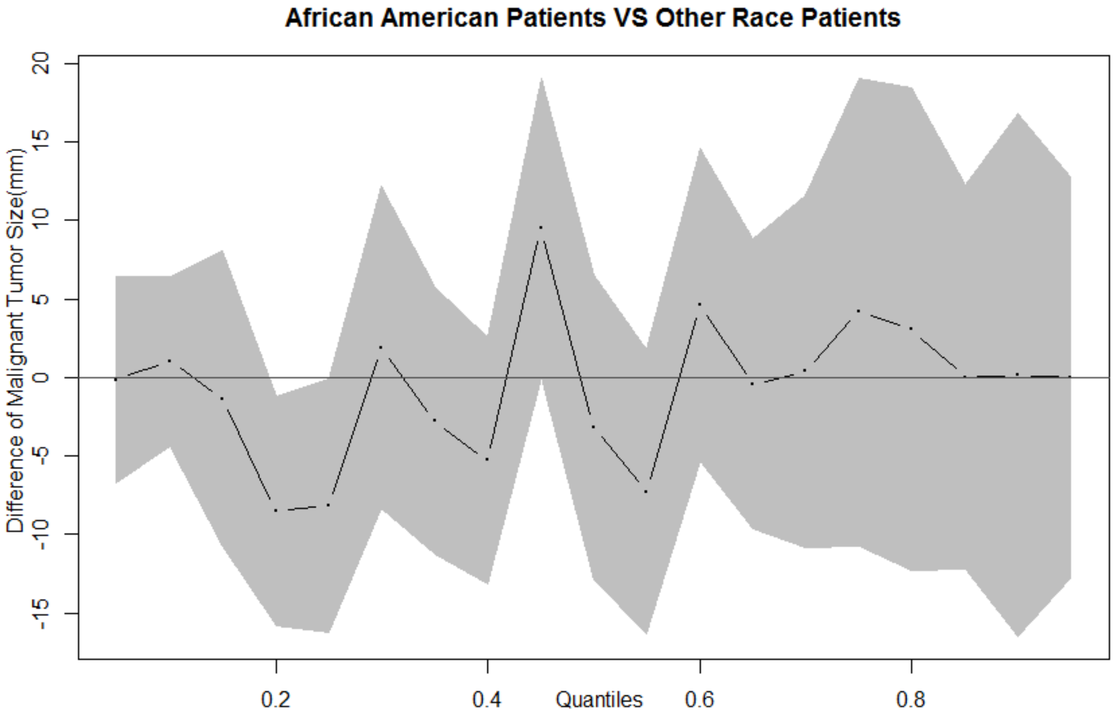


Figure 3.4: The estimated difference of malignant tumor size between African American and other race patients. The shaded area depicts at least 95% point-wise confidence band

Figure 3.5 below shows the estimated difference of malignant stomach tumor size between male and female patients, which matches the results obtained from Chapter 2. It proves that the malignant stomach tumor size for male patients is bigger than that for females. For example, for 50th quantile, the male patients will have the bigger malignant stomach tumor size than that of female and we have at least confidence to conclude that the difference will be 7 and 19 millimeters.

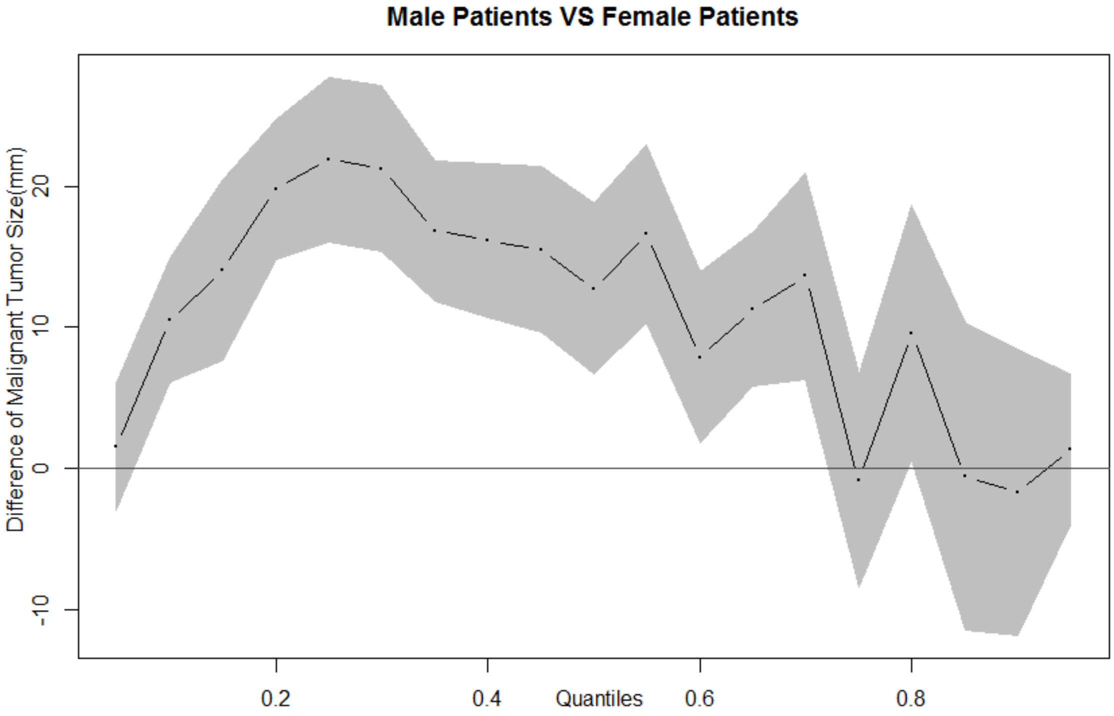


Figure 3.5: The estimated difference of malignant tumor size between male and female patients. The shaded area depicts at least 95% point-wise confidence band

The following equations supports the statistical visualization of the first quantile, median quantile and third quantile regression models with the appropriate estimates of the coefficients.

The 25th quantile:

$$Q_{0.25} = 15.5 - 13.6 * I(white) - 8.17 * I(black) +$$

$$21.9 * I(male) + 0.125 * age - 0.27 * age * I(male)$$

$$+0.14 * age * I(white) + 0.14 * age * I(black). \qquad (3.14)$$

The median quantile:

$$Q_{0.50} = 40 - 7.71 * I(white) - 3.16 * I(black) +$$

$$12.72 * I(male) + 0.003 * age - 0.136 * age * I(male)$$

$$+0.092 * age * I(white) + 0.08 * age * I(black) \qquad (3.15)$$

The 75th quantile:

$$Q_{0.75} = 60.83 - 0.83 * I(white) + 4.16 * I(black) -$$

$$0.83 * I(male) + 0.04 * age - 0.04 * age * I(male)$$

$$-0.04 * age * I(white) - 0.04 * age * I(black)$$

$$+3.83 * I(male) * I(white) \qquad (3.16)$$

Table 3.1:Ordinary and Quantile regression estimates: the coefficients on "malignant tumor size" are reported for 19 quantiles.

	Intercept	Race1	Race2	Sex 1	Age	Age*Sex1	Age*Race1	Age*Race2	Sex1*Race1	Sex1*Race2
OLS	**39.288**	-3.41	0.5630	**13.67**	**0.092**	**-0.186**	0.0027	0.0225	1.1503	0.6134
Quantile Regression from 5th to 95th quantiles:										
5	1.292	-2.12	-0.153	1.571	**0.097**	-0.011	-0.010	-0.0022	1.505	2.289
10	2.142	**-6.23**	1.035	**10.49**	**0.1428**	**-0.115**	0.0476	-0.0045	1.3214	-0.0928
15	**7.2490**	**-10.5**	-1.345	**14.025**	**0.1432**	**-0.168**	0.0875	-0.025	**2.969**	-0.070
20	**14.090**	**-14.7**	**-8.51**	**19.78**	**0.091**	**-0.2522**	**0.161**	**0.1612**	0.9002	-0.3607
25	**15.50**	**-13.6**	**-8.171**	**21.907**	**0.125**	**-0.2731**	**0.14815**	**0.148**	**1.2037**	**0.76389**
30	**16.875**	**-10.9**	1.90	**21.22**	**0.156**	**-0.2702**	**0.1140**	0.0114	**2.833**	-0.9966
35	**27.21**	**-13.1**	-2.788	**16.845**	0.042	**-0.198**	**0.156**	0.1006	-0.9722	-1.496
40	**28.5**	-5.92	-5.27	**16.15**	**0.0833**	**-0.1996**	0.0422	**0.1162**	0.7913	0.6221
45	**25.178**	-0.17	**9.494**	**15.52**	**0.1785**	**-0.1883**	-0.0357	-0.0101	1.795	-0.6403
50	**40.0**	**-7.71**	**-3.167**	**12.727**	**0**	**-0.136**	**0.0952**	**0.0833**	**-1.764**	**-0.1240**
55	**40.07**	-3.99	-7.272	**16.59**	0.086	**-0.219**	-0.0095	0.11355	2.186	1.852
60	**45.41**	**-6.30**	4.588	**7.890**	0.0588	**-0.1034**	0.0446	-0.0588	**3.006**	1.5233
65	**49.123**	0.876	-0.423	**11.26**	0.046	**-0.1538**	-0.0461	0.0538	0.738	0.4307
70	**51.4328**	3.567	0.3947	**13.619**	0.104	**-0.2051**	**-0.1044**	-0.001	1.4320	1.5859
75	**60.833**	**-0.83**	**4.1667**	**-0.833**	**0.0416**	**-0.0416**	**-0.0416**	**-0.0416**	**3.833**	**3.75**
80	**66.85**	-1.85	3.070	**9.578**	0.05	**-0.1785**	-0.05	-0.0479	2.7428	4.5012
85	**80**	-6.6	0	-0.583	0	-0.0833	-0.1667	0	**5.2833**	**5.9166**
90	**89.85**	-0.98	0.1428	-1.714	-0.071	0	0.0714	0.0714	1.714	1.714
95	**100**	**7.12**	0	1.374	0	-0.0218	**-0.16**	0	0.5963	0.0872

Table 3.1 above provides the estimated coefficients of our statistical quantile regression model from 5th to 95th quantiles. The results include the estimates coefficients for gender, race, age and their interaction terms. The red numbers denote the significant effect of

tumor size on that covariate at significant level 0.05. Table 3.1 shows that the variable age appears to be a statistically significant factor that affects the change in malignant tumor size for lower quantile regression models. When the other covariates remain at the same level, for each year increase in age, the statistically significant effects on the stomach tumor sizes are 0.097, 0.142, 0.143, 0.09, 0.125, 0.156, 0.083 and 0.178, respectively for the quantile regressions of 5th, 10th, 15th, 20th, 25th, 30th, 40th, 45th. For the higher quantiles, the patient's age effect of the conditional probability on the tumor size is not very significant. The standard ordinary least squares estimation shows that the stomach tumor size increases 0.092 when the age increases by one unit.

We also found that the interaction term between age and gender always indicates the significant effects in stomach malignant tumor size from 5th to 80th quantiles. We also identified that the covariate gender provides significant effect on the tumor size except the upper quantiles. Since covariate female is the controlling variable, estimated quantile regression coefficients of gender represent the difference of the malignant tumor size between male and female patients while the other factors age and race are fixed.

3.4 Discussion

Malignant Tumor size is strongly related to prognosis in the medical research area. In general, the smaller the tumor, the higher the chances are for long-term survival. The goal of this study is to find out the importance of the covariate age, gender and race on the effect of the malignant stomach tumor size. It is desirable to utilize the relationship between patients' demographic information and malignant stomach tumor size in order to help researchers postulate histology specific etiologic risk factors. We have applied both the statistical ordinary least squares and statistical quantile regression models to investigate the effects of the risk factors age, gender, race and their interaction terms on the malignant stomach tumor size.

The stomach tumor registry data was provided by SEER program. We demonstrated that patients' histology classification could lead the differences in the tumor size. We explored the statistical model of malignant stomach tumor sizes using statistical quantile regression models that assist to identify the effects of patients' demographic information associated with changing tumor size. For example, the probability distribution of stomach tumor sizes for male patients is always higher than that of female patients.
The estimates of coefficients for the probability distribution of malignant tumor size of African American patients are lower than those of other race patients for low quantiles, whereas the middle and higher quantiles are significantly higher. When the covariates gender and race remain fixed, the predictor age will show a positive effect on the malignant stomach tumor size for low quantiles.

3.5 Contributions

In this study, we have developed a statistical quantile regression model that include the effect of gender, age, race and their interaction terms. Having developed this statistical quantile regression model, we obtained the following useful information:

1. We demonstrated the effect of patients' gender, age, race on the malignant stomach tumor sizes, that will be helpful to the physicians to make their decisions.
2. Based on the results of our statistical quantile regression model, we found that the malignant stomach tumor size is different between male and female patients. For instance, for 50th quantile, the malignant tumor size of male patients is 12.7 mm larger than that of female patients when other covariates hold constant. Such information allows the medical physicians to treat female and male patients differently.
3. Our study reveals a significant difference of malignant tumor size for patients among different groups of races. For example, for 35th quantile, white patients' malignant tumor size is 13.1 mm smaller than other race patients' malignant tumor size. However, for 45th quantile of African American patients' malignant tumor size is 9.49 mm bigger than other race patients' malignant tumor size.
4. Moreover, our work shows that the patients' age is a significantly contributing factor for estimating the malignant stomach tumor size. For 25th quantile, the malignant tumor size increases 0.125 mm for one-year increase in patients' age.
5. We also have found that a significant interaction exists between age and gender.
6. The developed statistical quantile regression model provides a more flexible and comprehensive comparison of malignant stomach tumor size. Such information is essential for medical scientists to improve the treatments for their patients.

Chapter 4

Statistical Modeling of Malignant Stomach Tumor Size as Function of Age

4.1 Introduction

According to the American Cancer Society, stomach cancer is the fifth leading cause of cancer and the third leading cause of death from cancer. The exact causes of stomach cancer are still unknown. There is a total of 11,462 records of stomach patients that were diagnosed with non-malignant and malignant tumors in United States from 2004 to 2013. The data was obtained from the Surveillance, Epidemiology and End Results program (SEER).

Our previous study has shown that malignant stomach tumor sizes significantly differ on genders and races. The statistical quantile regression model shows that the patients' age is an additional significant variable on the size of malignant tumors.

In the present study, we aim to investigate the effect of age on stomach cancer tumor size for different types of genders and races.

4.2 White Male Malignant Stomach Tumor Size and Age

From the SEER database, there are 4,945 white male patients with malignant stomach tumors and about 83% of the cases were aged from 40 to 80 years. Our statistical models will focus on the majority part of the patients age. Figure 4.1 below shows a scatter plot of white male stomach patients with malignant tumor sizes in millimeters.

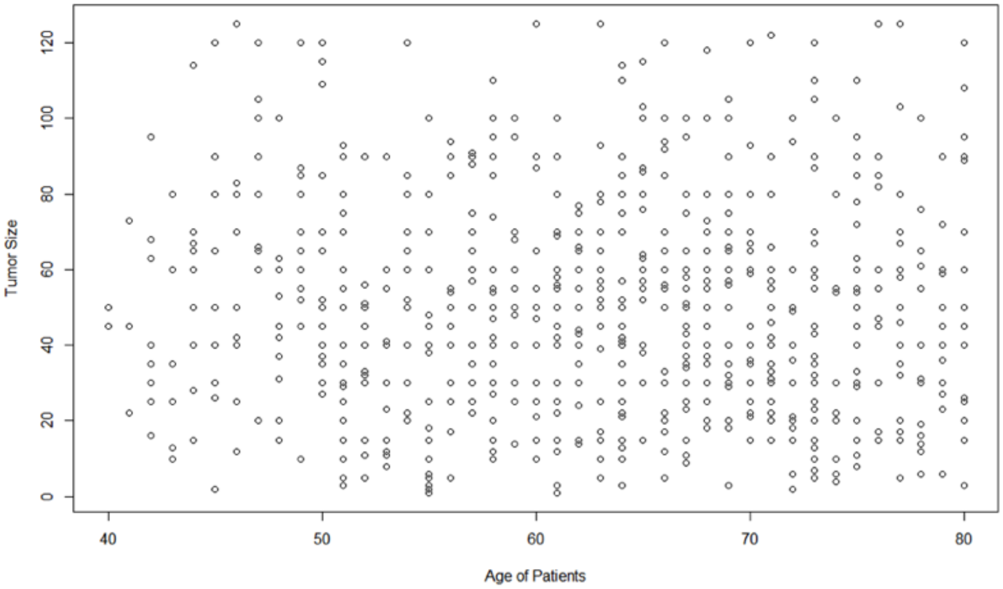


Figure 4.1: Raw data of White Male patients with stomach tumor size in millimeters

The raw data plot doesn't display any pattern. Therefore, we averaged the patients' malignant tumor size at each age level. Figure 4.2 below shows the scatter diagram of averaging malignant stomach cancer tumor sizes as a function of age for white male patients.

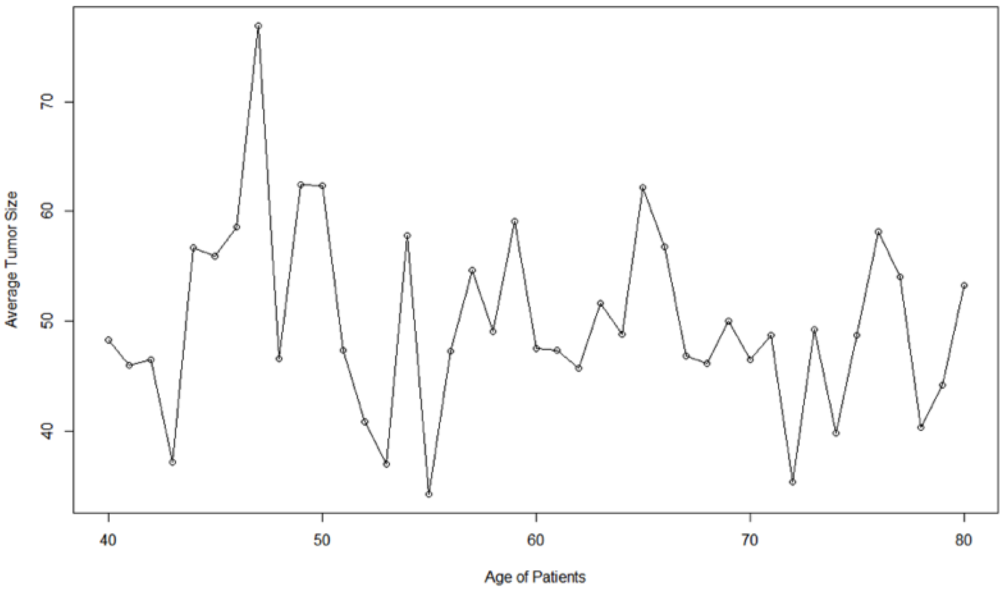


Figure 4.2: Mean of Malignant Stomach Tumor size for White Male in millimeters

Figure 4.2 above indicates that the curve has an inflexion point almost every three or four years of age, which makes it difficult to determine a function that characterize the mean behavior of malignant tumor sizes as a function of age (Kottabi, 2012). In order to address this issue, we averaged the malignant tumor size within each four-year interval. That is, we have 10 intervals of 4 years in length.

We denote stomach cancer patients' age with a. The corresponding malignant tumor size as a function of age is denoted by $T(a)$ in millimeters. The rate of the tumor size $T'(a)$ is the derivative of the function $T(a)$.

The best statistical model that characterizes white male cancer patients with stomach tumor size is given by the following equation:

$$T(a) = 1.49 * 10^6 + 2.35 * 10^3 a - 13.56 * a^2 + 3.84 * 10^{-2} a^3$$
$$-1.209 * 10^6 e^{\frac{1}{a}} - 8.908 * 10^4 \log(a), 40 \leq a \leq 80. \qquad (4.1)$$

The evaluation of the quality of the proposed model with respect to R^2, R^2 adjusted and residual analysis are given in Table 4.1 below:

Table 4.1: White Male Residual Analysis of Stomach Cancer Tumor Size

Sum of Residuals	-0.01018
Sum of Squared Residuals	0.69756
R-square	0.87
Adjusted- R square	0.84

Thus, the proposed statistical model shows a very good quality to predict the malignant tumor size as a function of age. Given below is a graph of the model along with an approximate 95% confidence limits. Thus, from Figure 4.3 we can obtain the following information such that if a white male patient, 60 years old, the approximate expected malignant tumor size will be about 44.26 millimeters and we are 95% confident that his tumor size will be between 43.18 and 45.34 millimeters. And on the other hand, if a white male patient has a 45 millimeters malignant tumor size then his age is almost 52 years old.

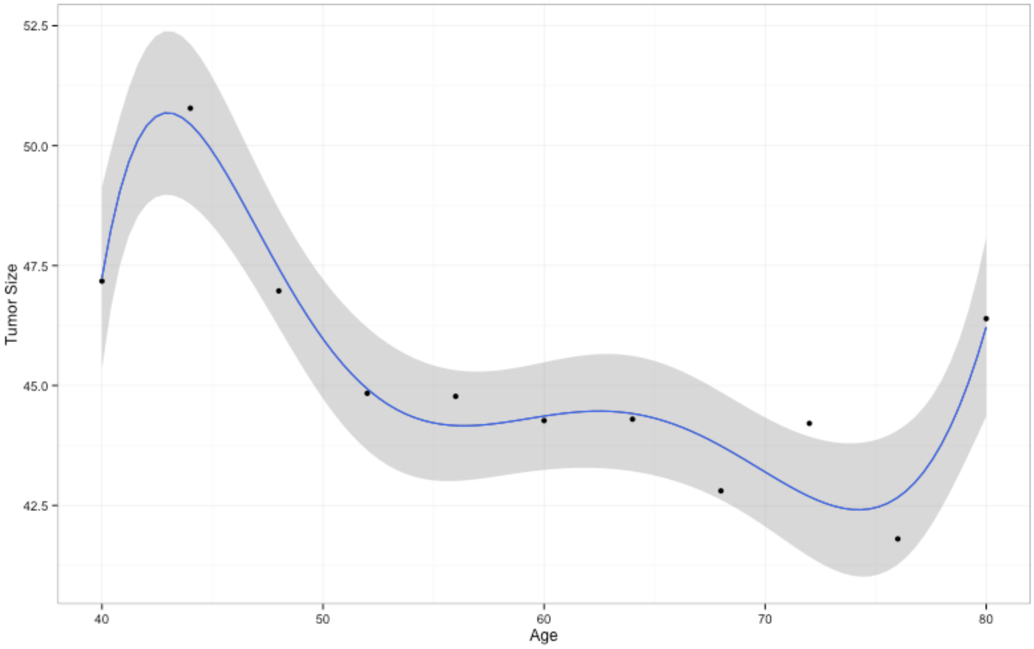


Figure 4.3: Estimated predicted model with a 95% confidence interval for white male in millimeters

The derivative of Equation 4.1 estimates a measure of the change of the mean malignant tumor size as a function of age. That is,

$$T'(a) = 2.35 * 10^3 - 27.12a + 11.52 * 10^{-2}a^2$$

$$+1.209 * \frac{10^6 e^{\frac{1}{a}}}{a^2} - 8.908 * \frac{10^4}{a}, 40 \le a \le 80 \qquad (4.2)$$

Thus, if one wants to find the rate of the malignant tumor size at a particular age, they can use the above function to predict the changing rate of the malignant tumor size. For instance, if a 61 years old patient, we can conclude that his changing rate of the malignant tumor size will be $T'(61) = 0.1047$ millimeters.

The classical rate of change (CRC) of mean malignant tumor sizes with respect to age is given by the following function (Bonsu, 2013, Kottabi, 2012, Chan, 2013)

$$CRC = \frac{T(a)\ in\ current\ age - T(a)\ in\ previous\ age}{T(a)\ in\ previous\ age} \tag{4.3}$$

We repeated some age of the white male stomach cancer patients and we calculated the tumor size, Equation 4.1, the rate of changing, Equation 4.2, and the classical rate of change, CRC, Equation 4.3, the results are given in Table 4.2 below:

Table 4.2: Residual Analysis of Rate Change of Mean tumor size for White male

Age	Tumor	Rate of Change(CRC)	Rate=$T'(a)$	Rate of Residual
48	46.97	-0.0285	-0.7066	0.6781
49	46.50	-0.01	-0.6978	0.6878
50	44.98	-0.0326	-0.655	0.6224
51	45.60	0.0136	-0.5871	0.6007
….	…………	………………………	………………..	…………………………
60	44.26	-0.0047	0.0934	-0.0981
61	44.19	-0.0017	0.1047	-0.1064
Mean of Residual Rate			-0.032	
Standard Deviation of Residual			0.6544	

Table 4.2 reveals that the residual is quite small as well as the standard deviation, which indicates that the proposed model is of good quality to characterize the behavior of the tumor size as a function of age.

Figure 4.4 below shows how the rate of the mean of the malignant tumor sizes changes as a function of the patients' age increases as one would expect. For instance, if a patient 60 years old we can conclude that his changing rate of the malignant tumor size will be $T'(61) = 0.0934$ millimeters.

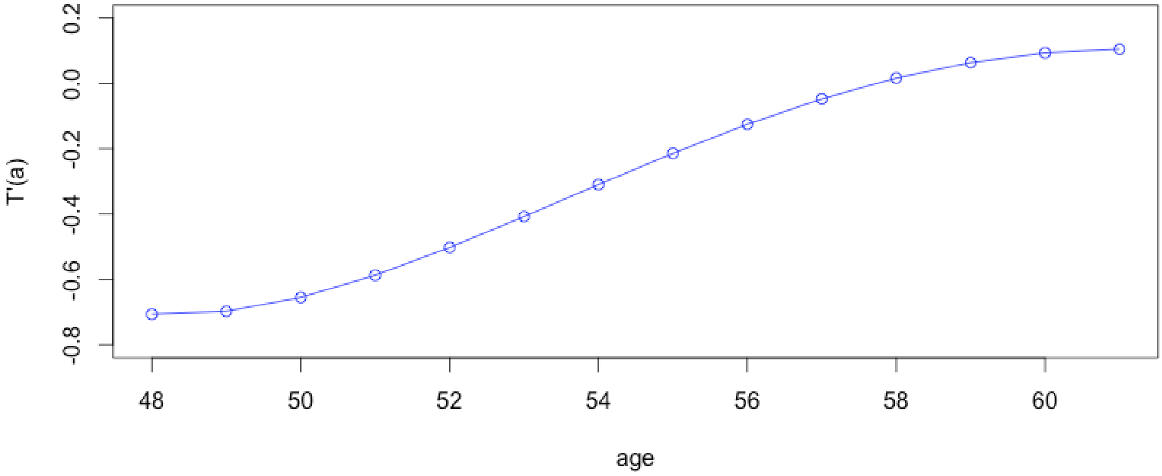


Figure 4.4: Changing Rate of stomach tumor size for white male

4.3 African American Male Malignant Stomach Tumor Size and Age

In this section, we will propose a statistical model of the African American Male stomach tumor size as a function of age. Figure 4.5 below shows that the curve has a turning point almost every three or four years of age, thus we focus on the behavior of the average size of the malignant tumors within a three-year interval.

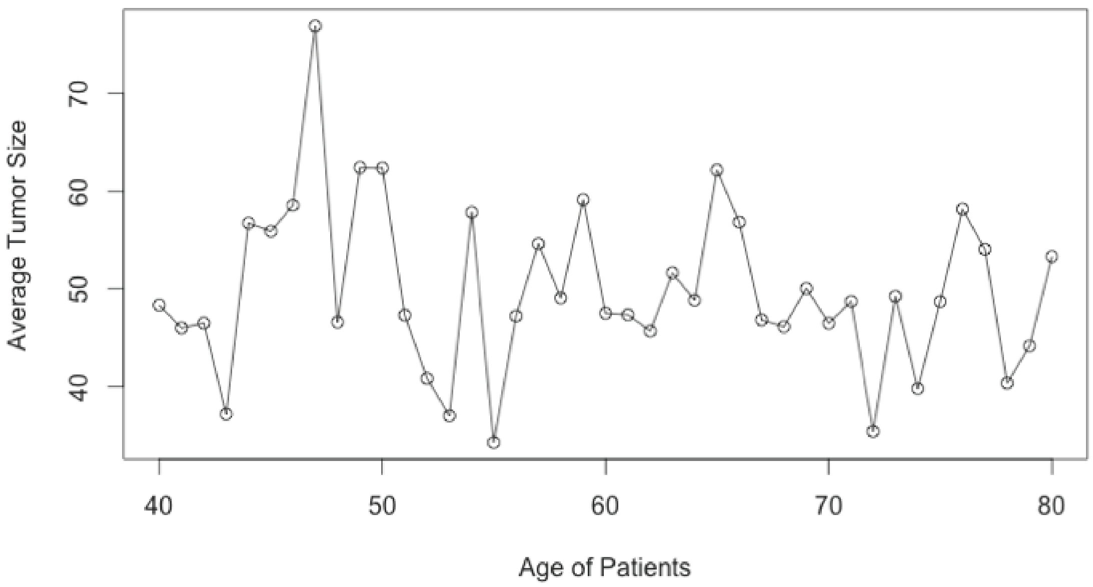


Figure 4.5: Mean of Malignant Stomach Tumor size for African American Male

After searching many kinds of nonlinear statistical model, we could not find an appropriate statistical model to describe the behavior of the malignant tumor size as a function of age for all the patients. Then we clustered the patients into two groups, 40 to 55 years old and older than 55 years old.

The statistical function that characterizes the stomach cancer tumor size as a function of age for African American male patients is expressed in the following Equation 4.4 for patients from 40 years old to 55 years old.

$$T(a) = 3.849 * 10^4 + 776.4a - 4.422a^2 - 1.692 * 10^4 \log(a)$$
$$+5.439 * 10^{-23} e^a, 40 \leq a \leq 55 \qquad (4.4)$$

This model that characterizes the malignant tumor size as a function of age for African American male patients from age 40 to 55 is a good model. It has an R^2 of 0.86 with R^2 adjusted of 0.82 and a very insignificant residual mean. A summary of this information is given in Table 4.3 below:

Table 4.3: African American Male Residual Analysis of Stomach Cancer Tumor Size from 40 to 55 years old

Sum of Residuals	0.00002
Sum of Squared Residuals	0.8252
R-square	0.86
Adjusted- R square	0.82

Figure 4.6 below shows the predicted curve of the mean malignant tumor size as a function of age along with at least 95% confidence interval for African American male patients from 40 to 55 years old. Thus, we can obtain the following information such that if an African American male patient, 50 years old, the approximate expected malignant tumor size will be about 55.81 millimeters and we are 95% confident that his tumor size will be between 51.62 and 60 millimeters. And on the other hand, if an African American male patient has a 50 millimeters malignant tumor size then his age is almost 43 years old.

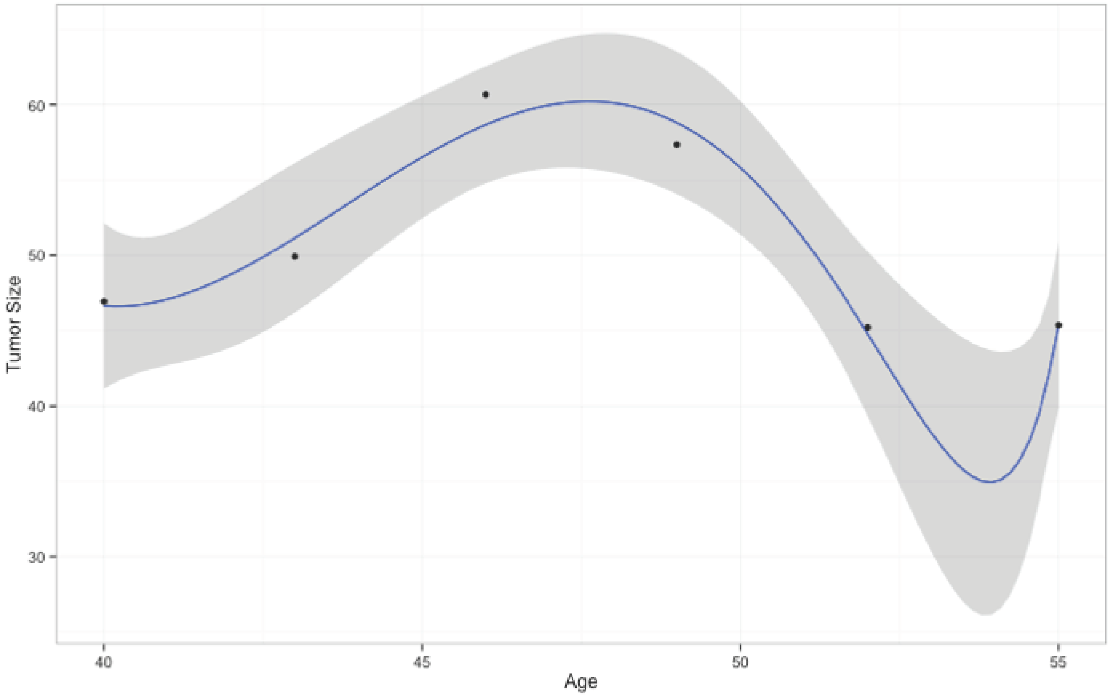


Figure 4.6: Estimated predicted model with a 95% confidence interval for African American male from 40 to 55

The derivative of Equation 4.4 estimates a measure of the change of the mean malignant tumor size as a function of age for African American male patients between 40 and 55 years old. That is,

$$T'(a) = 776.4 - 8.844a - \frac{16920}{a} + 5.439 * 10^{-23} e^{a}, 40 \leq a \leq 55. \qquad (4.5)$$

Thus, if one wants to estimate the changing rate of the malignant tumor size at a particular age, they can use the above function to predict the changing rate of the malignant tumor size. For example, if an African American male stomach cancer patient, 48 years old, we can conclude that his changing rate of the malignant tumor size will be $T'(48) = 0.746$ millimeters.

Next, we will use the similar process to obtain the statistical model respect to the malignant tumor size as a function of age for African American male older than 55 years old. That is,

$$T(a) = 1.324 * 10^7 + 3.75 * 10^5 a - 4.214 * 10^3 a^2 + 27.94a^3 \\ -6.26 * 10^6 \log(a) + 2.791 * 10^{-33} e^a, 55 < a \leq 80 \qquad (4.6)$$

This model that characterizes the malignant tumor size as a function of age for African American male patients older than 55 is a good model. It has an R^2 of 0.86 with R^2 adjusted of 0.83 and a very insignificant residual mean. A summary of this information is given in Table 4.4 below:

Table 4.4: African American Male Residual Analysis of Stomach Cancer Tumor Size older than 55 years

Sum of Residuals	0.00
Sum of Squared Residuals	1.153
R-square	0.86
Adjusted- R square	0.83

Thus, the proposed statistical model shows a very good quality to predict the malignant tumor size as a function of age for African American male patients older than 55 years.

Figure 4.7 below shows the predicted curve of the mean malignant tumor size as a function of age along with at least 95% confidence interval for African American male patients older than 55 years old.

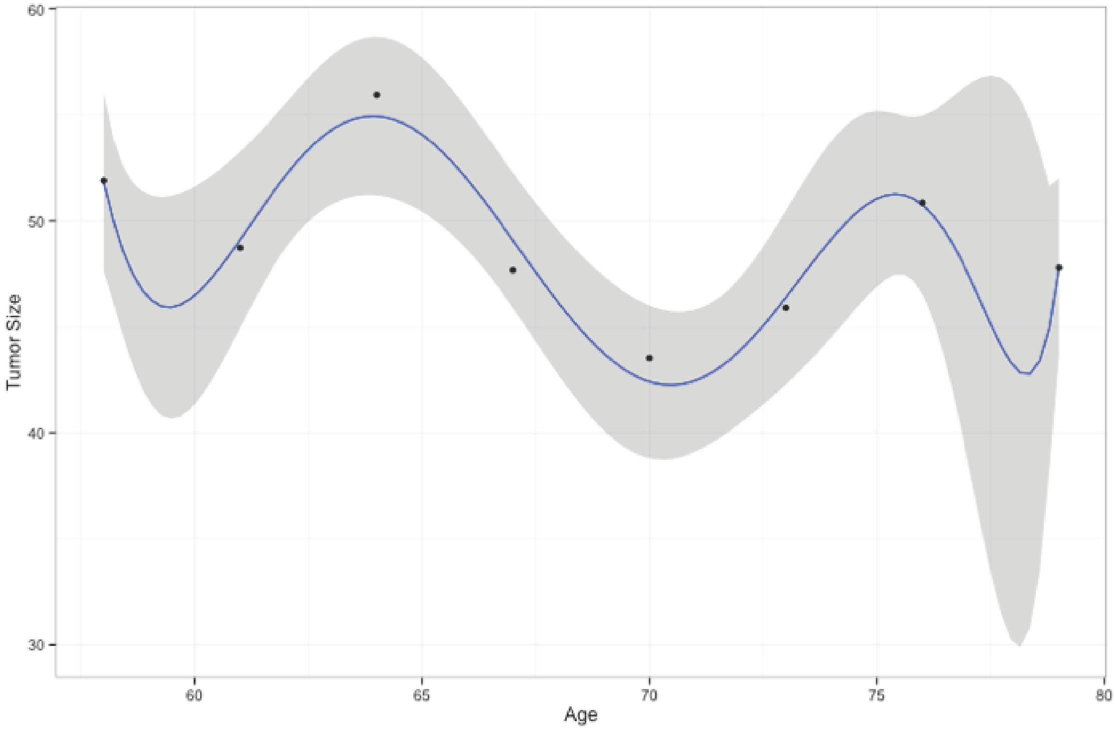


Figure 4.7: Estimated predicted model with a 95% confidence interval for African American male older than 55

Thus, we can obtain the following information such that if an African American male patient, 60 years old, the approximate expected malignant tumor size will be about 46.84 millimeters and we are 95% confident that his tumor size will be between 41.56 and 52.12 millimeters. And on the other hand, if an African American male patient has a 48.72 millimeters malignant tumor size then his age is almost 62 years old.

Then the derivative of Equation 4.6 indicates a measure of the change of the mean malignant tumor size as a function of age for African American male older than 55 years old. That is,

$$T'(a) = 3.75 * 10^5 - 8.4 * 10^3 a + 83.82a^2$$

$$-6.26 * \frac{10^6}{\mathrm{a}} + 2.791 * 10^{-33} e^a, 55 < a \leq 80 \qquad (4.7)$$

Thus, we can use the above function to predict the changing rate of the malignant tumor size for African American male patients older than 55 years old. For instance, if the stomach cancer patient 60 years old, we can conclude that his changing rate of the malignant tumor size will be $T'(60) = 0.2636$ millimeters.

In order to evaluate the accuracy of the proposed model, we calculated the classical rate of change of mean tumor size with respect to age by using Equation 4.3, then compared with the result from Equations 4.5 and 4.7. The results of the residual analysis are shown in Table 4.5 below:

Table 4.5: Residual Analysis of Rate Change of Mean tumor size for African American male

Age	Tumor	Rate of Change(CRC)	Rate=$T'(a)$	Rate of Residual
53	43.03	0.0086	-0.0979	0.1065
54	46.439	-0.0077	-0.074	0.0663
55	45.36	0.0016	-0.035	0.0366
56	50.30	-0.0177	0.0165	-0.0342
57	54.26	-0.0067	0.0769	0.0836
58	51.89	-0.0034	0.1419	0.1453
59	51.32	-0.0063	0.2061	-0.2124
60	46.84	0.0168	0.2636	-0.2468
61	48.22	0.0079	0.3075	0.2996
62	48.72	0.0022	0.3301	-0.3279
63	44.26	0.0057	0.3429	-0.3372
64	44.19	-0.0366	0.3864	-0.423
Mean of Residual Rate			-0.0703	
Standard Deviation of Residual			0.2302	

Table 4.5 reveals that the residual is small as well as the standard deviation, which indicates that the proposed model is of good quality to characterize the behavior of the tumor size as a function of age.

Figure 4.8 below depicts the trend of the changing rate of the malignant stomach tumor as a function of African American male patients' age increases as one would expect.

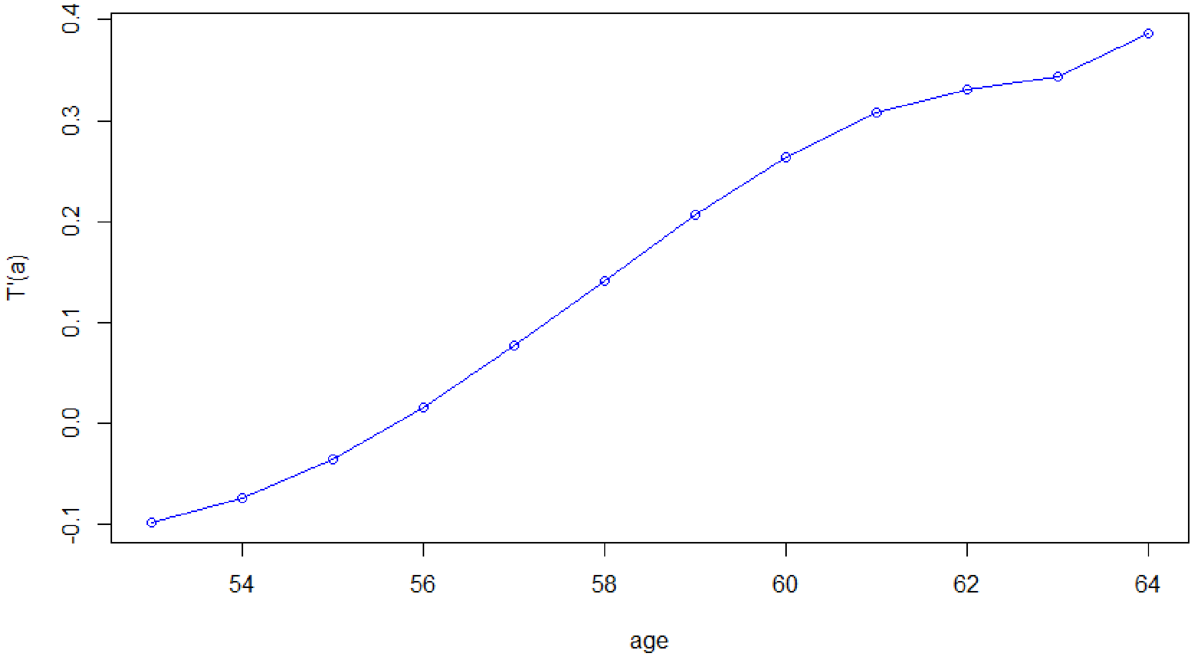


Figure 4.8: Changing Rate of stomach tumor size for African American male

4.4 Other Male Malignant Stomach Tumor Size and Age

In this section, we will propose a statistical model of all other male races stomach cancer tumor size as a function of age. Figure 4.9 below suggests that experiencing a change in the mean malignant tumor size every three or four years. Thus, we focus on taking the average size of the tumors within each three-year interval.

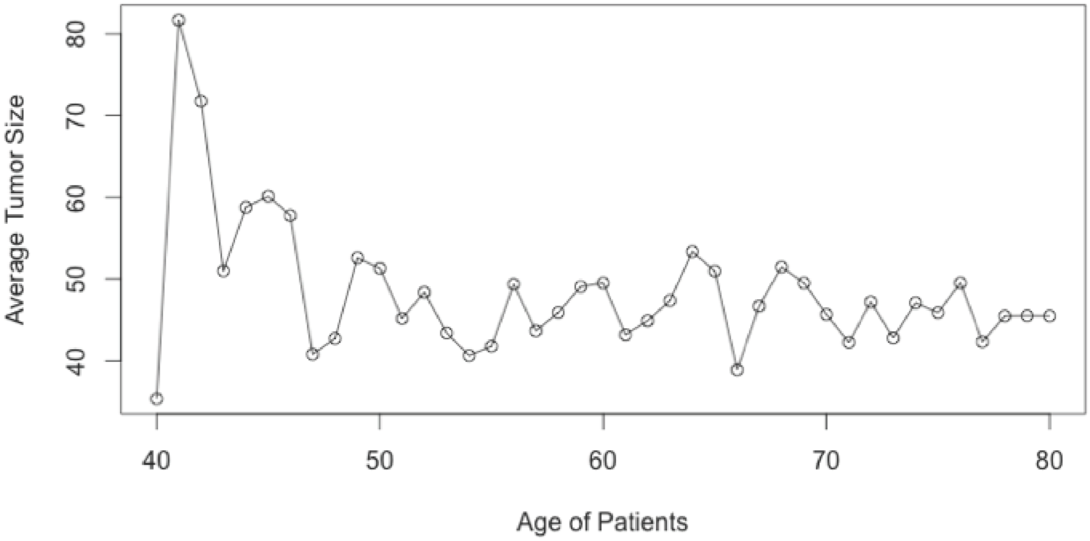


Figure 4.9: Mean of Malignant Stomach Tumor size for other Male

Then the nonlinear regression function that characterized the stomach cancer tumor size as the function of age for other male race patients is given by Equation 4.8:

$$T(a) = 2829 + 34.83a - 0.1405a^2 - 1066 * \log(a), 40 \leq a \leq 80. \quad (4.8)$$

The evaluation of the quality of the proposed model with respect to R^2, R^2 adjusted and residual analysis are given in Table 4.6, below:

Table 4.6: Other Race Male Residual Analysis of Stomach Cancer Tumor Size

Sum of Residuals	-2.7755e-17
Sum of Squared Residuals	0.7797
R-square	0.88
Adjusted- R square	0.86

Thus, the proposed model has a R^2 of 0.88, R^2 adjusted of 0.86, and the very small sum of residuals, which shows a very good quality to predict the malignant tumor size as a function of age.

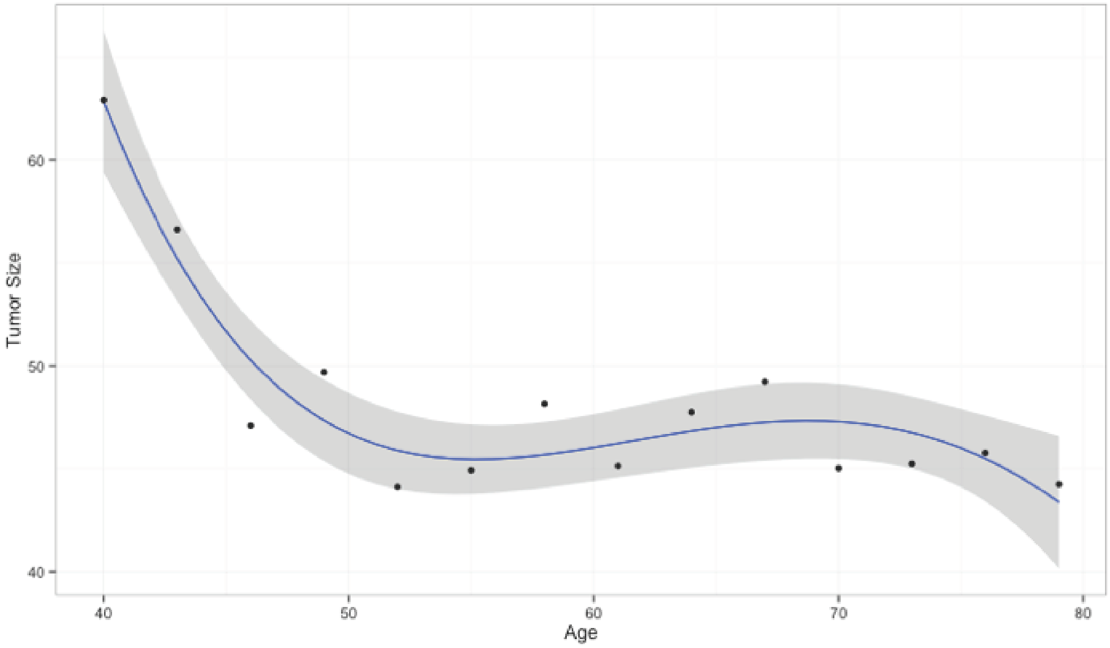


Figure 4.10: Estimated predicted model with 95% confidence interval for other male

Figure 4.10 above displays the predicted curve of mean malignant tumor size as a function of age with an approximate 95% confidence limits for other race male patients with stomach tumor. Thus, we can obtain the useful information such that a 55 years old other male race patient, the approximate expected malignant tumor size will be about 44.92 millimeters and we are 95% confident that his tumor size will be 42.1 and 47.74 millimeters. And on the other hand, if a patient has a 48 millimeters malignant tumor size then his age is almost 58 years old.

The following Equation 4.9 is the derivative function of other race male malignant tumor size as a function of age. That is,

$$T'(a) = 34.83 - 0.2810a - \frac{1066}{a}, 40 \leq a \leq 80. \qquad (4.9)$$

Thus, if the physician wants to find the changing rate of the malignant tumor size at a particular age of his patients, he can use the above function to predict the changing rate of his patients' malignant tumor size.

We calculated the tumor size, Equation 4.8, the rate of changing, Equation 4.9, and the classical rate of change, CRC, Equation 4.3, the results are given in Table 4.7 below:

Table 4.7: Residual Analysis of Rate Change of Mean tumor size for Other male

Age	Tumor	Rate of Change (CRC)	Rate=$T'(a)$	Rate of Residual
50	48.2718	-0.0284	-0.54	0.5115
51	45.6306	-0.0547	-0.4029	0.3482
52	44.1222	-0.0330	-0.282	0.2489
53	41.9277	-0.0497	-0.1762	0.1264
54	43.9192	0.0475	-0.0847	0.1322
55	44.9266	0.0229	-0.0068	0.0297
56	46.2984	0.0305	0.0582	-0.0277
57	46.1938	-0.0022	0.1112	-0.1135
58	48.1448	0.0422	0.1526	-0.1104
59	47.2388	-0.0188	0.1832	-0.2020
60	45.8488	-0.0294	0.2033	-0.2327
61	45.1314	-0.0156	0.2135	-0.2292
Mean of Residual Rate			0.07378	
Standard Deviation of Residual			0.2310	

Table 4.7 above reveals that the residual is small as well as the standard deviation, and such results indicate a good quality of the proposed model to characterize the behavior of the malignant tumor size as a function of age.

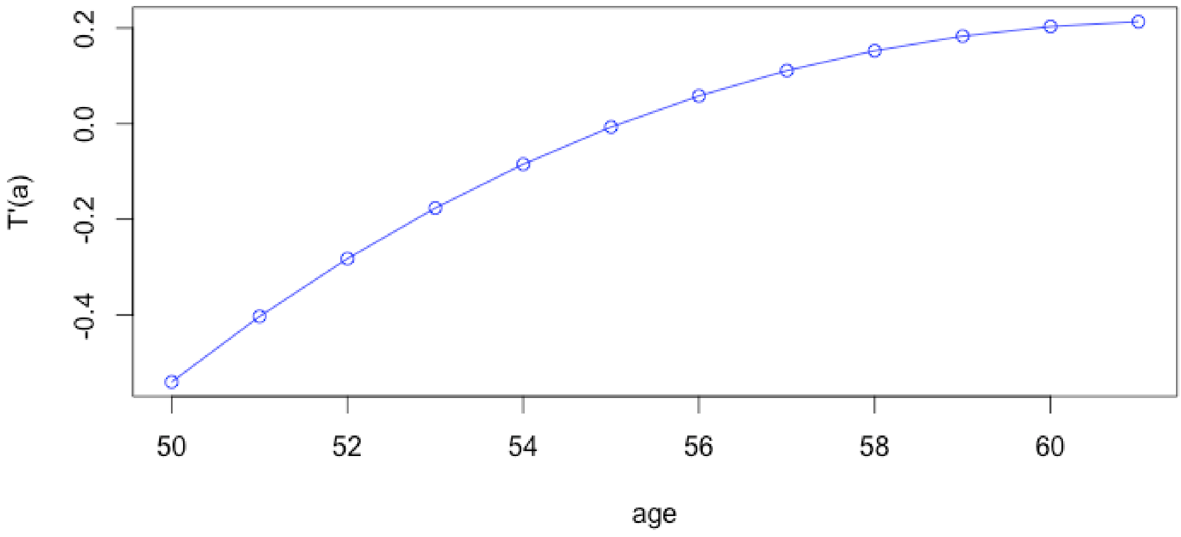


Figure 4.11: Changing Rate of stomach tumor size for other male

Figure 4.11 above depicts the trend of the changing size of the malignant stomach tumor for all other male race patients. For instance, if a 57 years old other male race patient, we can conclude that his changing rate of the malignant tumor size will be $T'(57) = 0.111$ millimeters.

4.5 White Female Malignant Stomach Tumor Size and Age

In this section, we will propose a statistical model of the white Female malignant stomach tumor size as a function of age. Figure 4.12 below assists us to find that the curve has an inflexion point almost every three or four years of age. Therefore, we focus on the average size of the malignant tumor within each three-year interval.

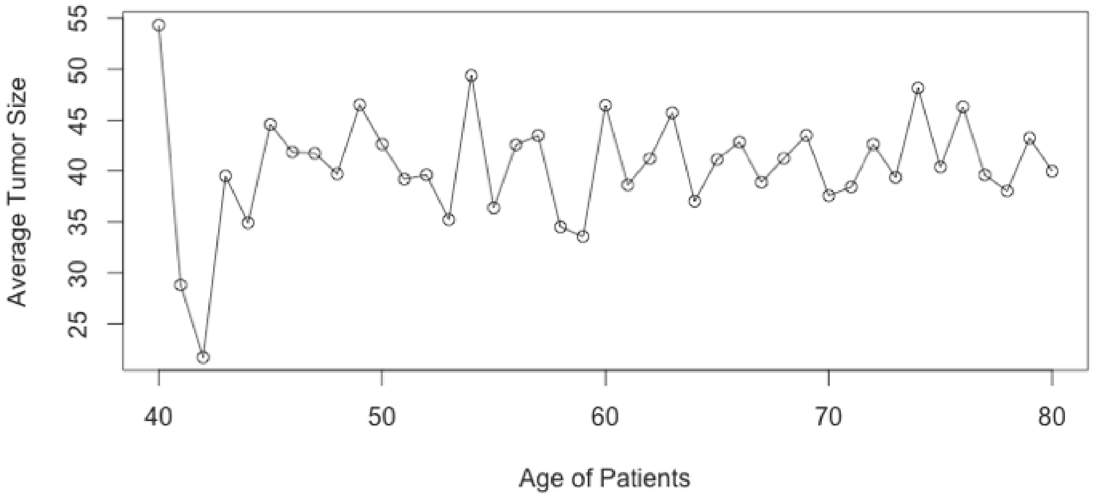


Figure 4.12: Mean of Malignant Stomach Tumor size for white female

The nonlinear regression function that characterizes the behavior of the stomach cancer tumor size as the function of age is given by the following Equation 4.10. That is,

$$T(a) = 1.535 * 10^5 + 111.9a - 6726\log(a)$$

$$-1.294 * 10^5 e^{\frac{1}{a}} - 0.303a^2, 40 \leq a \leq 80. \qquad (4.10)$$

The evaluation of the quality of the proposed model with respect to R^2, R^2 adjusted and residual analysis are given in Table 4.8, below:

Table 4.8: White Female Residual Analysis of Stomach Cancer Tumor Size

Sum of Residuals	0.0012e-25
Sum of Squared Residuals	0.949
R-square	0.80
Adjusted- R square	0.78

Thus, the R^2 is 0.80, the R^2 adjusted is 0.78 and the sum of residuals is small which indicates the proposed model shows a very good quality to predict the malignant tumor size as a function of age.

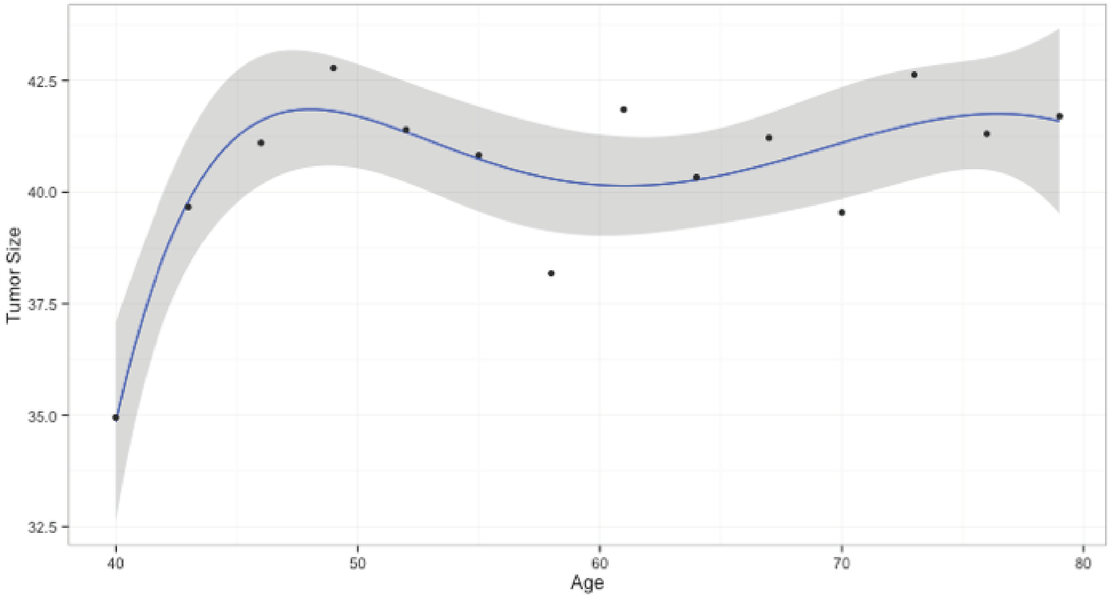


Figure 4.13: Estimated predicted model with a 95% confidence interval for white female

Figure 4.13 above is a graph of the model along with an approximate 95% confidence limits. Thus, we can obtain the important information such that a white female patient, 64 years old, the approximate expected malignant tumor size will be about 40.33 millimeters and we are 95% sure that her tumor size will be between 39.19 and 41.47 millimeters. And from the other side, if a white female patient with 41 millimeters malignant tumor size, then her age is almost 65 years old.

The following Equation 4.11 measures the change of the mean malignant tumor size as a function of age. That is,

$$T'(a) = 111.9 - \frac{6726}{\mathrm{a}} + 1.294 * \frac{10^5}{a^2 e^{\frac{1}{a}}} - 0.606a, 40 \leq a \leq 80. \qquad (4.11)$$

In order to evaluate the accuracy of our proposed model, we calculated the tumor size, Equation 4.10, the rate of change of mean tumor size, Equation 4.11, and the classical rate of change, CRC, Equation 4.3, and the results of the residual mean are given in Table 4.9 below:

Table 4.9: Residual Analysis of Rate Change of Mean tumor size for White female

Age	Tumor	Rate of Change	Rate=$T'(a)$	Rate of Residual

53	40.321	-0.0260	-0.1800	0.15398
54	42.776	0.0608	-0.1742	0.23515
55	40.823	-0.0456	-0.1591	0.11350
56	40.201	-0.0152	-0.1369	0.12170
57	37.194	-0.0747	-0.1094	0.03467
58	38.176	0.0263	-0.0784	0.10483
59	39.548	0.0359	-0.0453	0.08128
60	42.094	0.0643	-0.0114	0.07583
61	41.852	-0.0057	0.0220	-0.02783
62	41.318	-0.0127	0.0543	-0.06708
63	41.286	-0.0007	0.0843	-0.08516
64	40.331	-0.0231	0.1115	-0.13468
65	40.952	0.0154	0.1351	-0.11970
66	40.991	0.0009	0.1545	-0.15364
67	41.211	0.0053	0.1694	-0.16404
68	40.771	-0.0106	0.1792	-0.18993
69	39.833	-0.0229	0.1836	0.20666
Mean of Residual Rate			-0.0134	
Standard Deviation of Residual			0.1375	

In Table 4.9 above, the residual is small as well as the standard deviation, which indicates that the proposed model is of good quality to characterize the behavior of the malignant tumor size as a function of age.

Figure 4.14 below shows how the rate of the mean of the malignant tumor sizes changes as a function of the patients' age increases as one would expect. For instance, if a white female patient, 61 years old, we can conclude that her changing rate of the malignant tumor size will be $T'(61) = 0.022$ millimeters.

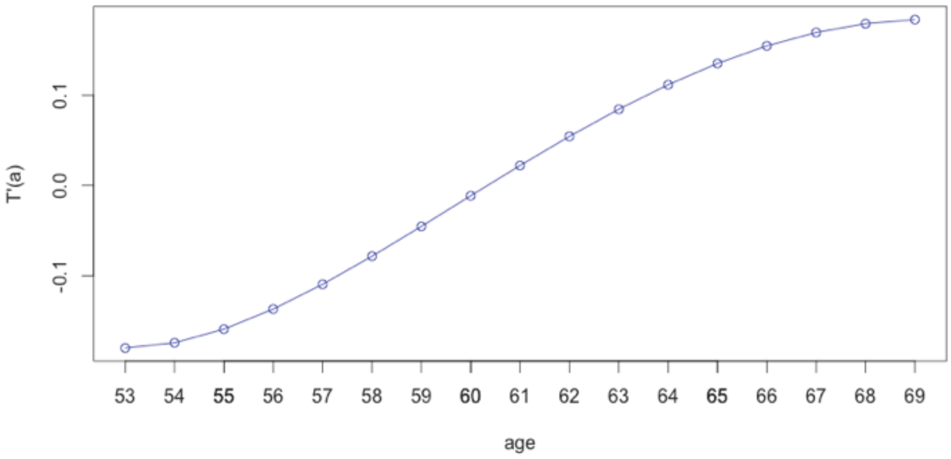


Figure 4.14: Changing Rate of stomach tumor size for white female

4.6 African American Female Malignant Stomach Tumor Size and Age

In this section, we will propose a statistical model of African American female stomach tumor size as a function of age. Figure 4.15 below points out that the curve has a change in the mean malignant tumor size almost every three or four years. Thus, we focus on the average size of the tumor within each three-years interval.

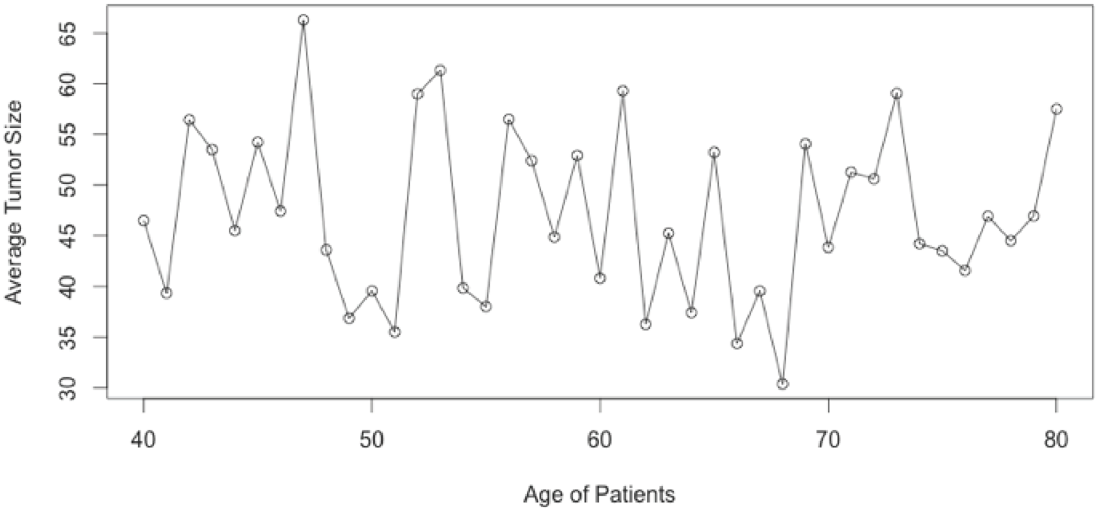


Figure 4.15: Mean of Malignant Stomach Tumor size for African American female

The best fitted nonlinear regression function that characterized the stomach cancer tumor size as a function of age for African American female patients with stomach tumor is expressed in the following Equation 4.12:

$$T(a) = -2821 - 41.84a + 1151 * \log(a) + 0.1846a^2, 40 \leq a \leq 80 \qquad (4.12)$$

This model that characterizes the malignant tumor size as a function of age for African American female patients is a good model. It has an R^2 of 0.90 with R^2 adjusted of 0.87 and a very insignificant residual mean. A summary of this information is given in Table 4.10 below:

Table 4.10: African American Female Residual Analysis of Stomach Cancer Tumor Size

Sum of Residuals	1.587e-17
Sum of Squared Residuals	0.768
R-square	0.90
Adjusted- R square	0.87

Figure 4.16 below displays the predicted statistical regression curve of mean malignant tumor size as a function of age with an approximate 95% confidence interval for African American female patients with malignant stomach tumor. Thus, we could obtain the following information such that if an African American female patient, 43 years old, the approximate expected malignant tumor size will be about 51.12 millimeters and we are 95% confident to conclude that her tumor size will be between 50.08 and 52.16 millimeters. And on the other side, if an African American female with 54 millimeters, then her age is almost 79 years old.

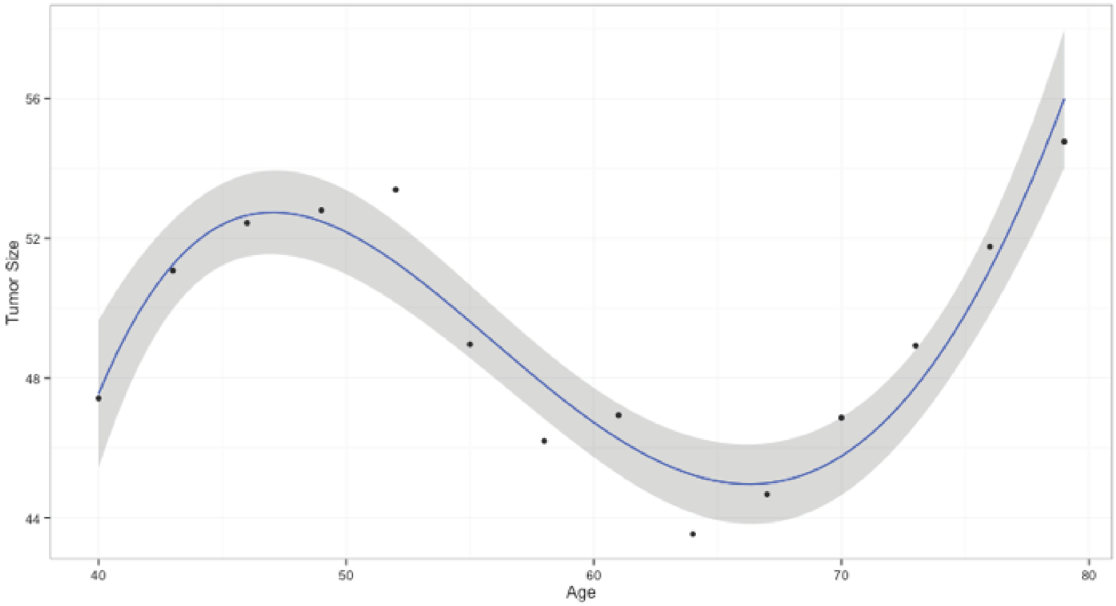


Figure 4.16: Estimated predicted model with 95% confidence interval for African American female

The following Equation 4.13 measures the change of the mean malignant tumor size as a function of age. That is,

$$T'(a) = -41.84 + \frac{1151}{a} + 0.3692a, 40 \leq a \leq 80. \tag{4.13}$$

Thus, we can use the above function to predict the changing rate of the malignant tumor size. For example, if a 68 years old patient, we can conclude that her changing rate of the malignant tumor size will be $T'(68) = 0.192$ millimeters.

Table 4.11: Residual Analysis of Rate Change of Mean tumor size for African American female

Age	Tumor	Rate of Change	Rate=$T'(a)$	Rate of Residual

56	51.255	0.0468	-0.611	0.658
57	50.071	-0.023	-0.602	0.579
58	46.199	-0.077	-0.581	0.504
59	51.007	0.104	-0.548	0.652
60	45.454	-0.108	-0.504	0.395
61	46.937	0.032	-0.449	0.482
62	42.874	-0.155	-0.385	0.229
63	45.299	0.1422	-0.310	0.453
64	41.670	-0.080	-0.226	0.146
65	42.381	0.017	-0.134	0.151
66	44.75	-0.179	-0.033	-0.146
67	41.325	0.188	0.075	0.113
68	42.761	0.034	0.192	-0.157
69	49.727	0.162	0.315	-0.153
70	48.577	-0.023	0.446	-0.469
71	53.645	-0.104	0.584	-0.480
72	51.295	-0.043	0.728	-0.772
73	48.925	-0.046	0.878	-0.924
74	43.092	-0.119	1.034	-1.154
75	43.995	0.020	1.196	-1.175
76	44.328	0.007	1.363	-1.356
77	46.121	0.040	1.536	-1.496
78	49.647	0.076	1.714	-1.637
79	54.763	0.103	1.896	-1.793
80	54.552	-0.003	2.083	-2.087
Mean of Residual Rate			-0.077	
Standard Deviation of Residual			0.8566	

In order to evaluate the accuracy of the proposed model, we calculated the classical rate of change of mean tumor size, CRC, by using equation 4.3, then compared it with the rate of change, Equation 4.13. The results of the residual analysis are displayed in Table 4.11 above. Table 4.11 reveals that the residual is quite small as well as the standard error. Such results indicate a good quality of the proposed model for the size of malignant tumor as a function of age.

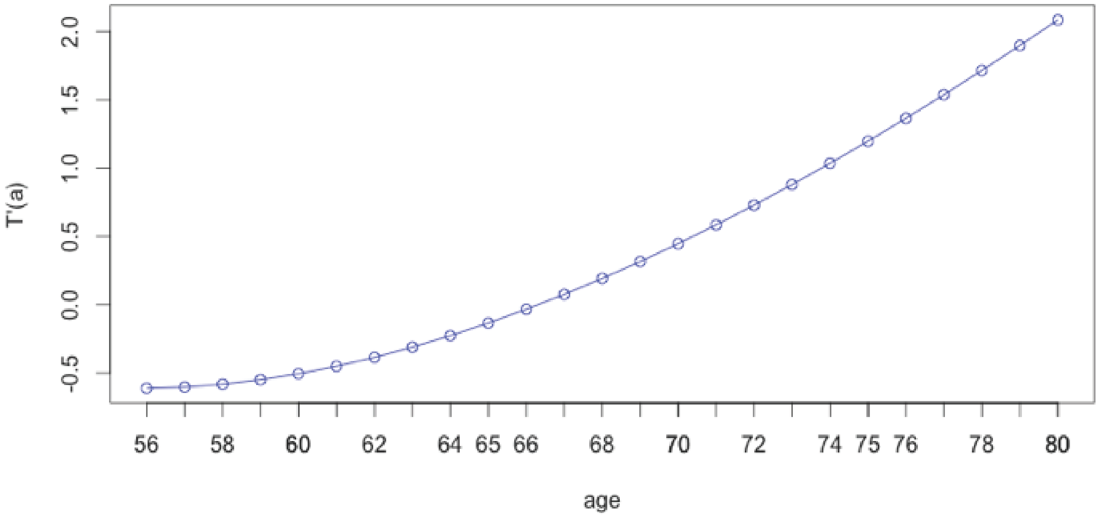


Figure 4.17: Changing Rate of stomach tumor size for African American female

Figure 4.17 above shows how the rate of the mean of the malignant tumor sizes changes as a function of the patients' age increases as one would expect. For instance, a 70 years old African American female, we can conclude that her changing rate of the malignant tumor size will be $T'(70) = 0.446$ millimeters.

4.7 Other Female Malignant Stomach Tumor Size and Age

In this section, we will propose a statistical model of all other race female malignant stomach tumor size as a function of age. Figure 4.18 below shows that the curve has a turning point for about every three or four years of age, and such result shifts our attention to average size of the tumor within each three-years interval.

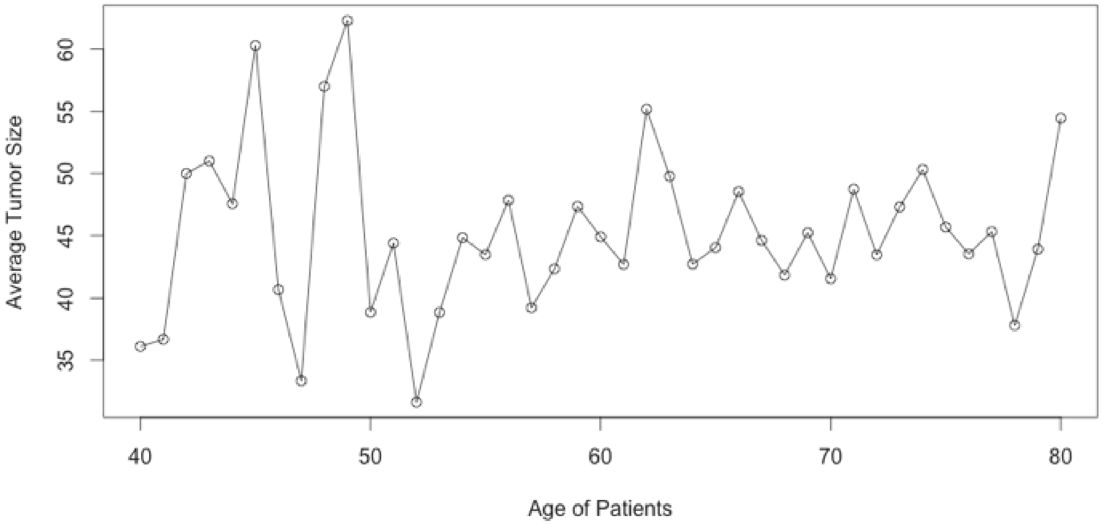


Figure 4.18: Mean of Malignant Stomach Tumor size for other female

The nonlinear regression function that characterized the malignant stomach cancer tumor size as a function of age for other female patients is given by the following Equation 4.14:

$$T(a) = 29540 + 139.9a - 0.3870a^2 - 8262 * \log(a)$$

$$-1.594 * \frac{10^5}{a}, 40 \leq a \leq 80. \qquad (4.14)$$

This model that characterizes the malignant tumor size as a function of age for other female race patients is a good model. It has an $R^2 = 0.84$ with R^2 adjusted of 0.80 and a very small residual mean. A summary of this information is given in Table 4.12 below.

Table 4.12: Other Race Female Residual Analysis of Stomach Cancer Tumor Size

Sum of Residuals	-4.5598e-17
Sum of Squared Residuals	0.5796
R-square	0.84
Adjusted- R square	0.80

Figure 4.19 below displays the predicted statistical regression curve of the mean malignant tumor size as a function of age with an approximate 95% confidence interval for other race female patients. For instance, if a 55 years old stomach cancer patient, the approximate expected malignant tumor size will be about 46.51 millimeters and we are 95% confident that her malignant tumor size will be between 45.78 and 47.24 millimeters.

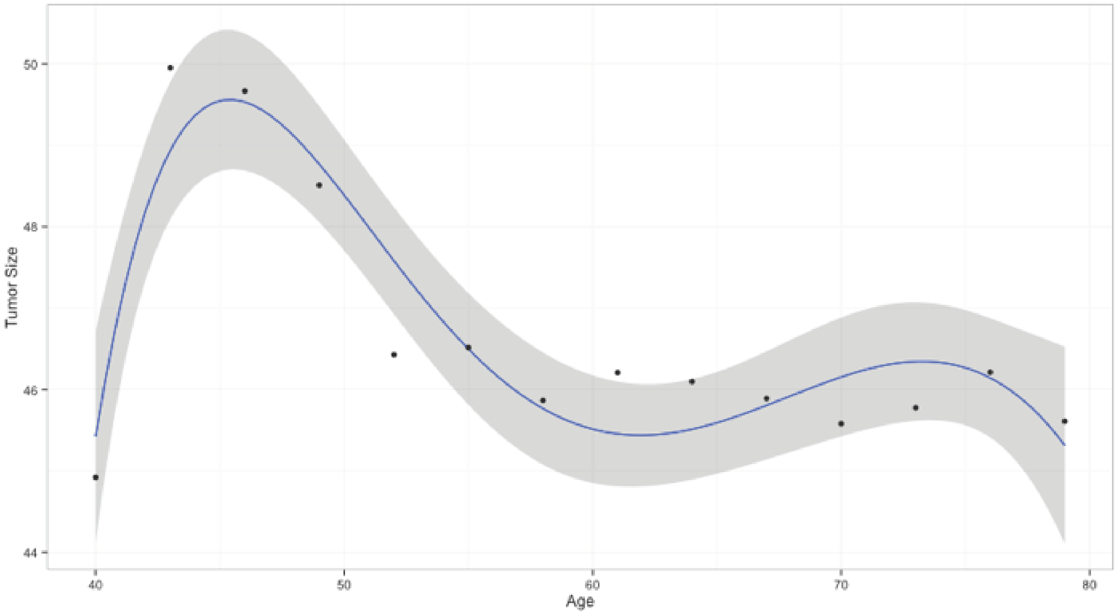


Figure 4.19: Estimated predicted model with 95% confidence interval for other female

The following Equation 4.15 shows the change of the mean malignant tumor size as a function of age. That is,

$$T'(a) = 139.904 - 0.7740a - \frac{8262}{a} + 1.594 * \frac{10^5}{a^2}, 40 \leq a \leq 80. \qquad (4.15)$$

In order to evaluate the accuracy of the results, we calculated the classical rate of change of mean tumor size with respect to age by using equation 4.3, then compared with the result from equation 4.15. The results of the residual analysis are displayed in Table 4.13 below.

Table 4.13: Residual Analysis of Rate Change of Mean tumor size for Other female

Age	Tumor	Rate of Change (CRC)	Rate=$T'(a)$	Rate of Residual
51	46.282	-0.0001	-0.2858	0.28577
52	46.427	0.0031	-0.2789	0.28201
53	44.380	-0.0440	-0.2586	0.21461
54	45.388	0.0227	-0.2280	0.25077
55	46.514	0.0248	-0.1899	0.21476
56	46.133	-0.0081	-0.1466	0.13853
57	45.967	-0.003	-0.1001	0.09711
58	45.866	-0.0022	-0.0522	0.05000
59	44.982	-0.0193	-0.0044	-0.01484
60	47.585	0.0578	0.0417	0.01602
61	46.208	-0.0289	0.0853	-0.11422
62	46.216	0.0001	0.1251	-0.12505
63	45.507	-0.0153	0.1603	-0.17569
64	46.098	0.0130	0.1902	-0.17726
65	45.733	-0.0079	0.2141	-0.22201
66	44.995	-0.0170	0.2313	-0.24831
67	45.889	0.0198	0.2415	-0.22178
68	45.867	-0.0004	0.2443	-0.24831
Mean of Residual Rate			0.00011	
Standard Deviation of Residual			0.19714	

Table 4.13 reveals that the residual is quite small as well as the standard error. The results indicate that the proposed model has a good quality to describe the behavior of the mean malignant tumor size as a function of age. Figure 4.20 below shows the trend of the changing size of the stomach tumor as a function of patients' age increases as one would expect. For instance, if a 64 years old patient, we can conclude that her changing rate of the malignant tumor size will be $T'(64) = 0.1902$ millimeters.

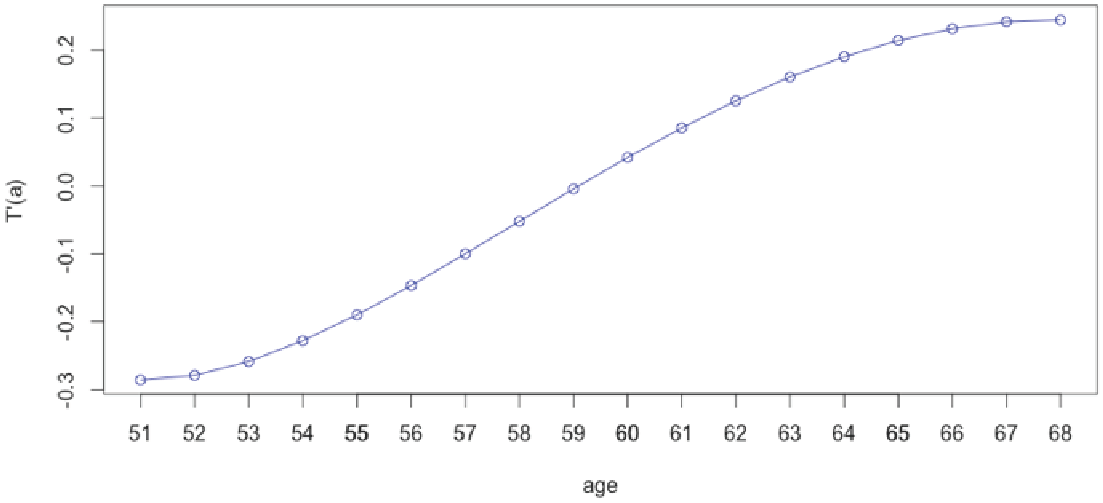


Figure 4.20: Changing Rate of stomach tumor size for other female

4.8 Contributions

In this chapter, we have developed the statistical models for the malignant stomach tumor size as a function of patient's age for different types of race and gender, respectively.

- Statistical models of stomach cancer tumor size for different groups of race and gender are quite different, and these models consist of linear and nonlinear functions.
- We are able to estimate the rate of change in stomach cancer tumor size based on the proposed statistical models.
- We have shown that the rate of mean stomach cancer tumor size, $T'(a)$, grows faster as the patients' age increases based on all the proposed statistical models.
- Both of the residual analysis of the statistical models indicate that we have found high-quality models of stomach cancer tumor size as the function of patients' age. Such outstanding results could assist the stomach physicians to make more accurate decisions.

Chapter 5

Stomach Cancer Treatment Effectiveness

5.1 Introduction

There are not so many studies that are related to the stomach cancer treatment effects. But we know that some studies can be found related to whether radiation or surgery shows a good effect to other cancer patients like breast cancer patients with respect to relapse time (Cong, 2010). However, the side effects including fatigue, mild skin reactions, upset stomach, and loose bowel movements from radiation therapy or surgery make it desirable to avoid surgery or radiation unless it is necessary. Therefore, it is very important for the physicians to identify the patients who could get benefit from surgery or radiation and those who could be at the very risk level to receive those treatments. In our study, we perform the nonparametric, parametric, and decision tree survival analysis to address this important question. Our parametric and nonparametric analysis identified that the overall advantage of combined radiation and surgery over radiation only in respect to the probability of survival times. With the utilization of the decision tree analysis in conjunction with survival analysis of survival time of stomach cancer patients, we have concluded that the subgroups of the two treatment groups affect the decision-making process in choosing the suitable treatment for stomach cancer patients.

5.2 Stomach Cancer Data

From 2004 to 2013, a total of 2786 patients were diagnosed, of which 2004 are male and 782 are female patients. For male patients, 595 patients received radiation only and 1409 received both radiation and surgery. For female patients, 183 patients received radiation only and 599 received both treatments.

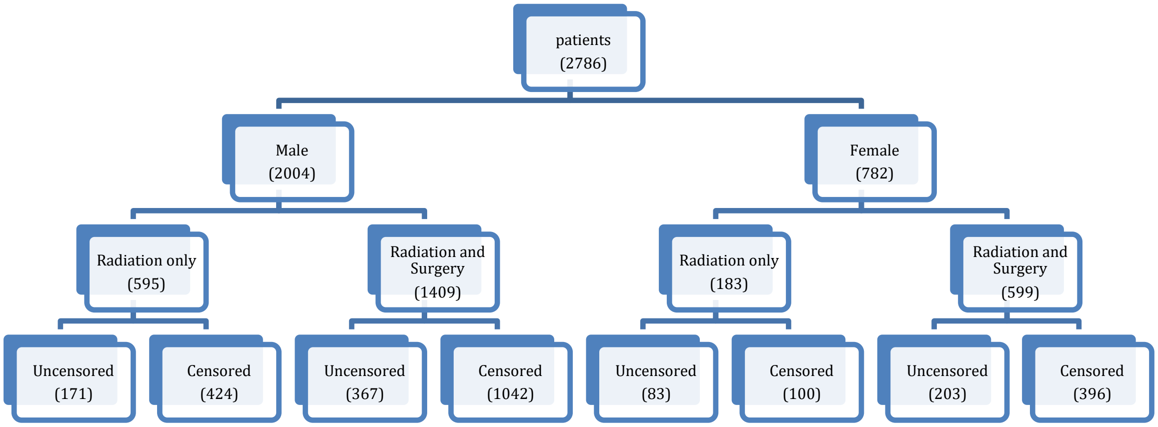


Figure 5.1: Patient treatments data

This censored data consists of 538 uncensored observations for male patients and 286 uncensored observations for female patients. On the other hand, it has 1466 censored observations for male and 496 censored observations for female patients as shown in Figure 5.1. The censored survival times are most likely due to two reasons: 1. The stomach cancer patients moved out of the study area; 2. The individual survived after the end of the study period. For male patients, 595 patients had the radiation alone and 1409 took a combination of radiation and surgery. For female patients, 183 took radiation only and 599 took both radiation and surgery. Since nearly 70% of the data are censored observations, we take into consideration two datasets for later analysis.

5.3 Nonparametric Survival Analysis

The Kaplan-Meier estimator is probably the most popular approach. We can use the empirical survival function: when there is no censoring, the general formula is:

$$\hat{S}(t) = \frac{\#\ \text{individuals with } T > t}{\text{total sample size}} = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t).$$

If there is censoring, the method is based on the ideas of conditional probability.so that

$$\hat{S}(t) = \prod_{r_j<t}\left(1 - \frac{d_j}{r_j}\right),$$

where d_j is the number of deaths at time t_j and r_j is the number of risk at time t_j. We can use the most common method Greenwood's formula to calculate the variance of the KM estimator:

$$\text{var}\left(\hat{S}(t)\right) = (\hat{S}(t))^2 \sum_{r_j<t} \frac{d_j}{(r_j - d_j)r_j}.$$

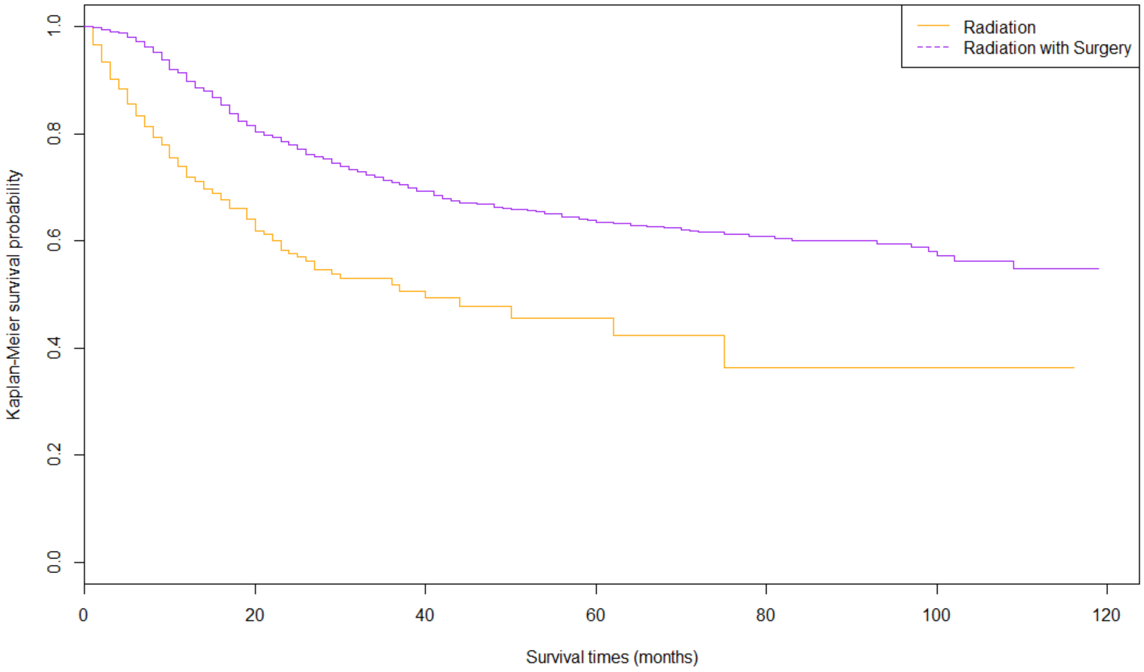


Figure 5.2: Survival curves of two treatments for male patients

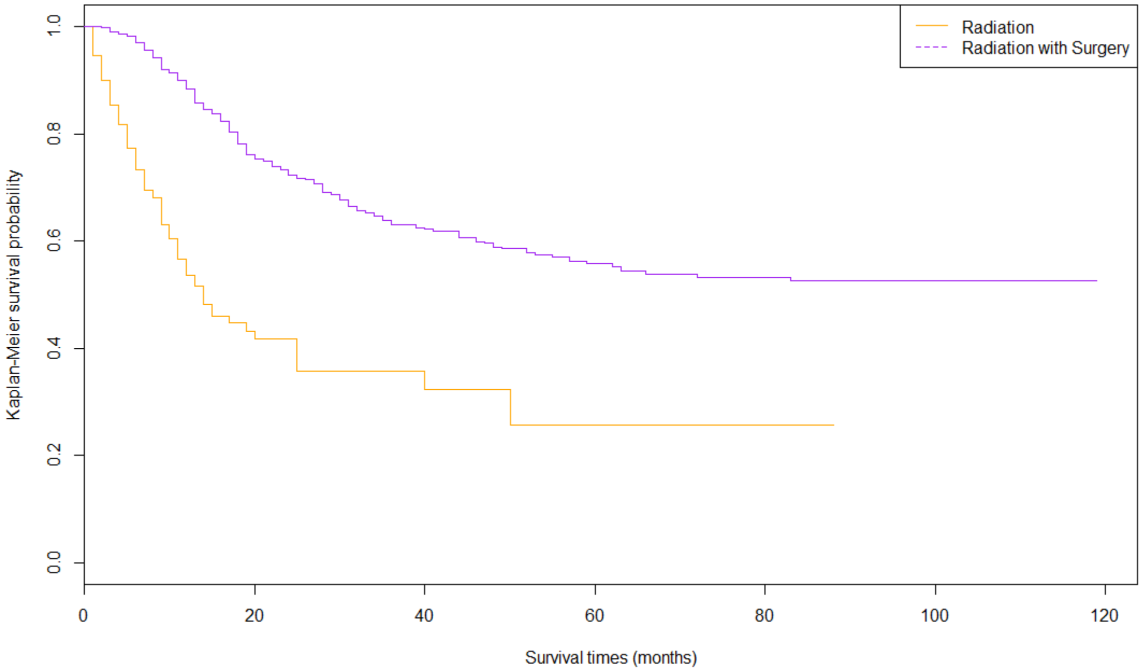


Figure 5.3: Survival curves of two treatments for female patients

Kaplan-Meier estimates of the survival curves of survival time for the two treatment groups for male and female patients are shown in Figure 5.2 and 5.3, respectively. In Figure 5.2 and 5.3, the two curves show they are significantly different and do not cross each other. Thus, it shows the patients who have surgery and radiation always have the better survival rate than that of patients who only receive radiation treatment.

An important advantage of the Kaplan–Meier curve is that it can take into account some types of censored dataset, which exists if a patient withdraws from a study, is lost to follow-up, or is alive without event occurrence at last follow-up. On the plot, small vertical tick-marks indicate individual patients whose survival times have been right-censored. When no truncation or censoring occurs, the Kaplan–Meier curve is the complement of the empirical distribution function. The Kaplan-Meier curve is the most common method to describe survival characteristics. The probability of surviving to any point is estimated from the cumulative probability of surviving each of the preceding time intervals. Although the probability calculated at any given interval is not very accurate because of the small number of events, the overall probability of surviving to each point is more accurate. However, Kaplan-Meier plot is not commonly used to compare the true mean effectiveness of the two treatments. In the present study, we perform actual nonparametric analysis utilizing Wilcoxon (Wilcoxon, 1963) rank sum test and Peto & Peto (Peto, 1972) modification of the Gehan-Wilcoxon test (Gross and Clark, 1975). By utilizing the two different nonparametric tests, we found the information in Table 5.1 below, which shows that the combination of the two treatments (radiation and surgery) is more effective than using the single treatment (radiation) which is consistent with Figure 5.2 and 5.3.

Table 5.1: Test the difference of means of two treatments for male and female patients

	Male	Female	Male	Female	Male	Female
	Chi-Square		Degree of freedom		P-value	
Log-Rank	92	96.5	1	1	7.963e-22	8.805e-23
Peto & Peto	104	111	1	1	1.967e-24	6.130e-26

5.4 Parametric Survival Analysis

In a parametric model, we assume the distribution of the survival curve to be known and the model are specified. Then the hazard function and the effect of any covariates can be obtained. First, we analyzed the censored dataset and found the generalized gamma distribution can be best characterized the behavior of survival time for male and female patients in different treatment groups, and the corresponding maximum likelihood estimator (MLE) is shown in Table 5.2. A graphical presentation of the cumulative distribution function (CDF) for male and female patients in radiation treatment group are shown by Figure 5.4 and 5.5 where the Kaplan-Meier curve and its 95% confidence band, as well as CDF of the fitted generalized gamma distribution, are plotted.

Table 5.2: Estimated parameters and log-likelihood of Generalized Gamma Distribution

	Male	Female	Male	Female	Male	Female	Male	Female
	mu		Sigma		Q		Log-Likelihood	
Total	3.977	3.482	2.139	1.932	0.850	0.970	-2981	-1507.9
Radiation	2.893	2.271	2.002	1.545	1.083	0.862	-805.5	-340.8
Radiation and Surgery	3.851	3.277	1.897	1.562	1.374	1.812	-2108	-1109.5

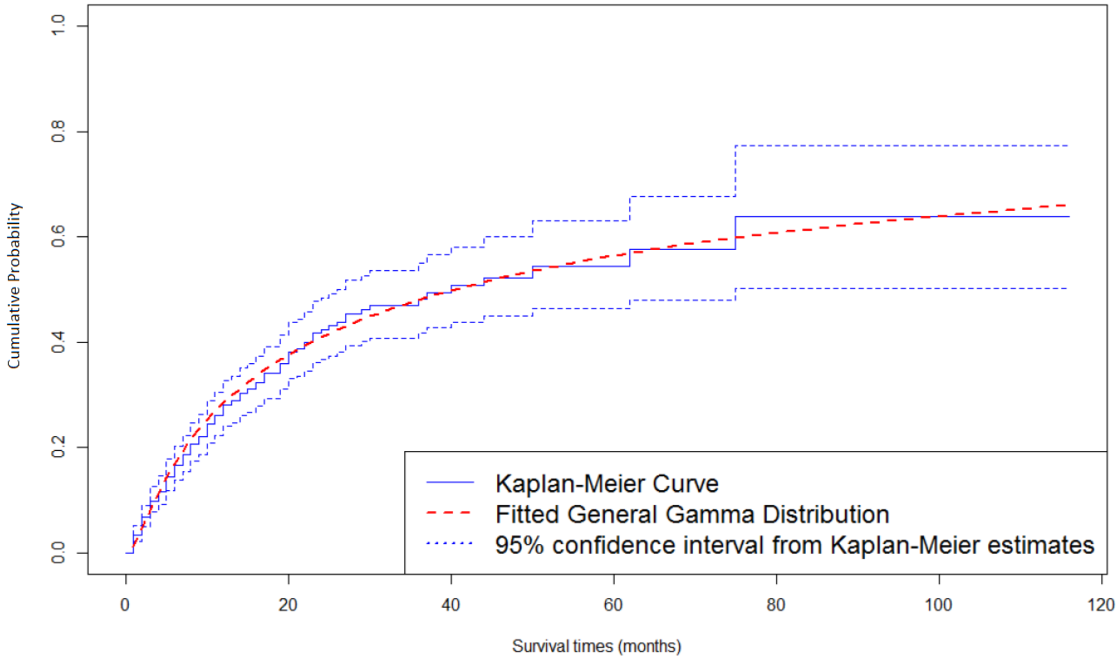


Figure 5.4: Fitted General Gamma Distribution of Radiation group for male patients

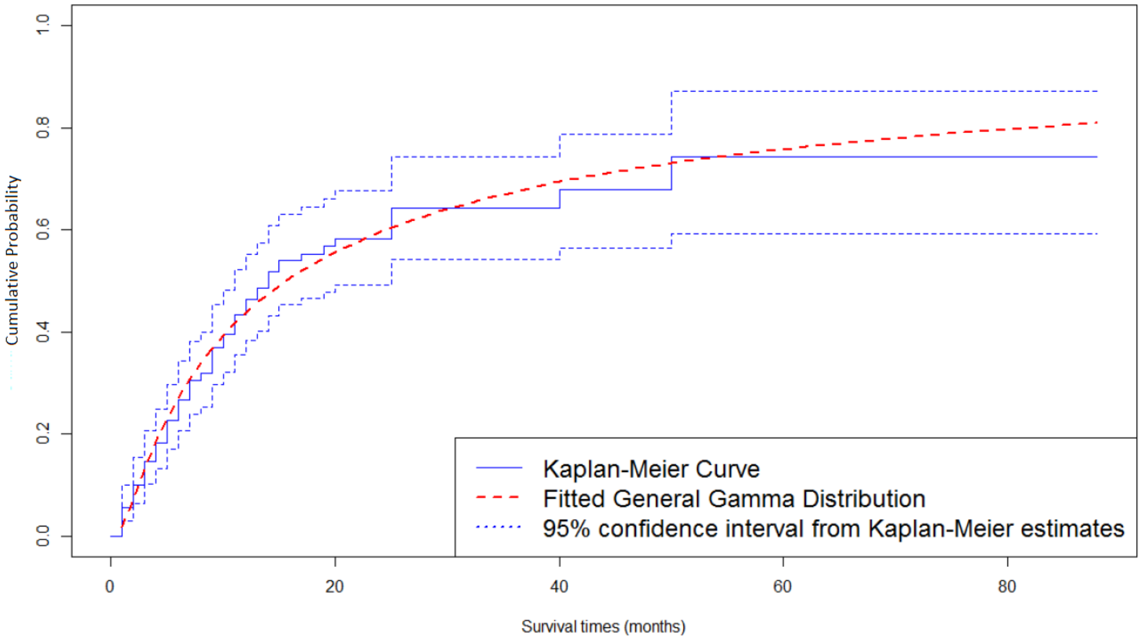


Figure 5.5: Fitted General Gamma Distribution of Radiation group for female patients

From Figures 5.4 and 5.5 above, the generalized gamma probability distribution seems to be a good fit for the survival time of male and female stomach cancer patients in the radiation treatment group. The fitted survival plots of the generalized gamma probability distribution are very close to the Kaplan-Meier survival plots and they are inside the 95% confidence limits of Kaplan-Meier survival plots.

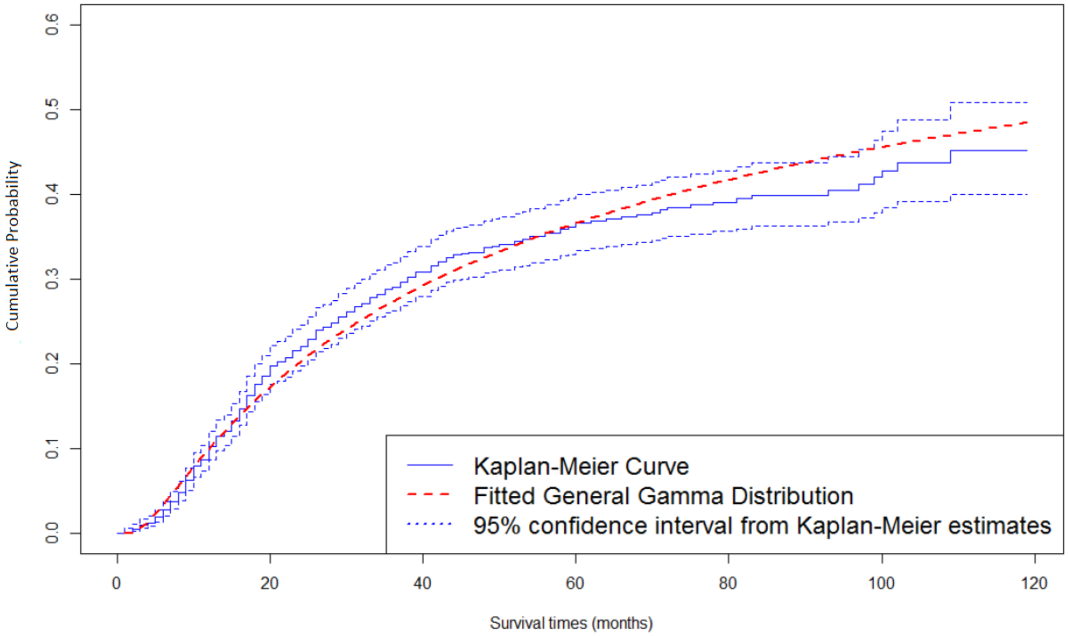


Figure 5.6: Fitted Distribution curve of Radiation and Surgery group for male

Similarly, we perform the same parametric analysis for male and female patients in the combination of radiation and surgery treatment group, and based on the goodness-of-fit test results, we have identified the generalized gamma distribution is the best fitted probability density function for the patients in both radiation and surgery treatment group. The corresponding maximum likelihood estimators are given in Table 5.2 and the plots of the cumulative distribution function are shown in Figure 5.6 and 5.7.

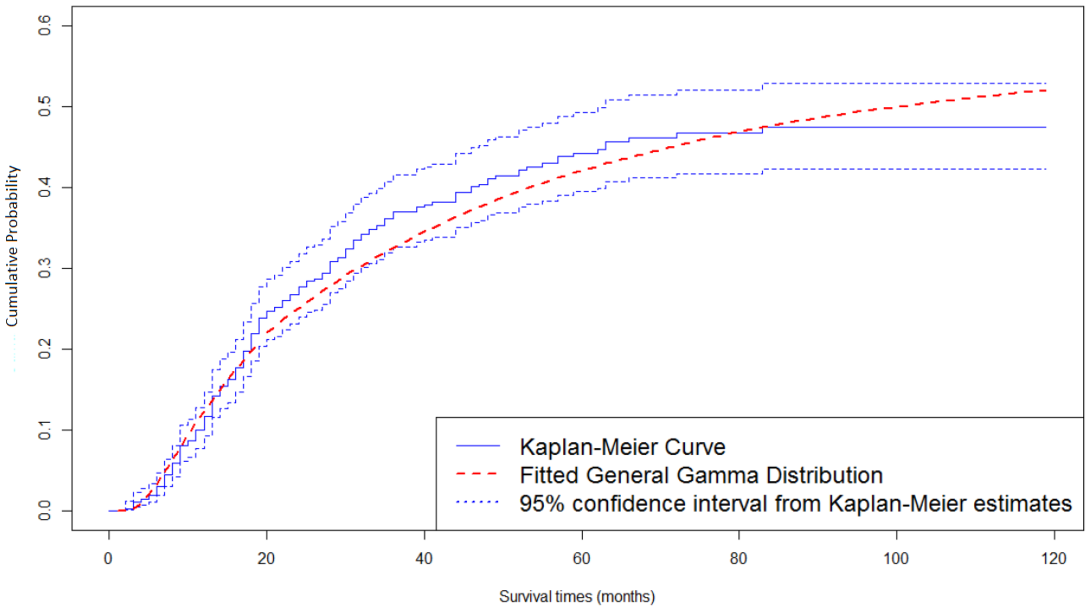


Figure 5.7: Fitted Distribution curve of Radiation and Surgery group for female

From Table 5.2, we know the survival time of the two treatment groups are both from the generalized gamma probability distribution. Then the log-likelihood ratio test could be performed to test the hypothesis, that is,

$$H_0: \mu_1 = \mu_2 = \mu \ \ vs. \ \ H_1: \mu_1 \neq \mu_2.$$

And the log-likelihood ratio test statistics is given below, That is,

$$T = -2[l(\mu,\mu) - l(\mu_1 - \mu_2)].$$

After the calculation, we found the test statistics are $T_{1,df=1} = 14.07$ (male) and $T_{2,df=1} = 12.27$ (female), and from the Chi-square distribution we found that the p-value are 0.0001 and 0.0004 for male and female patients respectively. Therefore, we can conclude that there is a significant difference between the two treatment groups for male and female patients, which is consistent with the previous conclusion when using the nonparametric test.

On the other side, for the uncensored dataset, we have 171 male patients in radiation treatment only and 367 male patients in radiation and surgery group. For female patients, we have 83 patients with radiation only and 203 patients are treated with the combination of radiation and surgery. Through goodness-of-fit tests which included Chi-Square, Kolmogorov-Smirnov and Anderson-Darling tests, we have identified that the best fitted parametric distribution function are general Pareto probability distribution for male patients with radiation only, log-logistic probability distribution for male patients with both radiation and surgery, Weibull distribution for female patients treated with radiation only and lognormal probability distribution for female patients in both radiation and surgery group. We have already identified the consistent results from the parametric and nonparametric test for censored database. Then we were only considering the censored data for further analysis.

5.5 Decision Tree Analysis

We are applying the decision tree analysis to partition the subject data as a function of the malignant tumor size and age of the patient. Decision tree analysis can be used to homogenize the information by separating the database into several different sub-groups based on similarity of survival time to treatment. Decision trees are helpful, not only because they are graphics that help you "visualize" what you are thinking, but also because making a decision tree requires a systematic, documented thought process. For instance, survival decision analysis provide the natural identification of predictive groups among stomach cancer patients, and the representation tree plots can help physicians to make early decisions regarding the treatments.

We performed the exponential decision tree analysis (Bacchetti, 1995) to reduce the impurity within nodes by partitioning based on the risk factors using a specified loss function. We denote the hazard rate within a given node as $h(y) = \lambda_j$ for all y in group j, thus the survival function within each node will be an exponential function. The loss function for the node t is shown by the following function, that is,

$$R(t) = -\hat{L}(t) = D_t - D_t \log\left(\frac{D_t}{Y_T}\right)$$

Where $D_t = \sum_i d_i$ is the number of complete observations and $Y_t = \sum_i y_i$ is the total observed event time.

Our goal is to compare the two different treatments together instead of single treatment only. We let the maximum tree depth to 3 and the complexity parameter 0.02. The trees of radiation only and both radiation and surgery treatment for male are shown in Figure 5.8 and Figure 5.9, respectively.

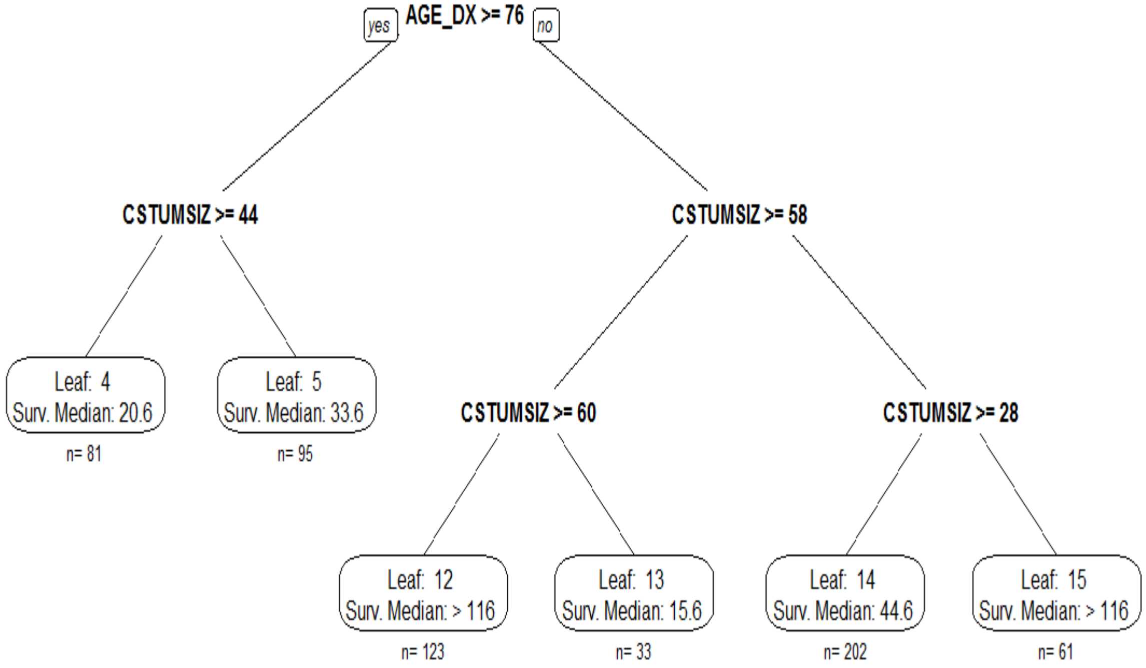


Figure 5.8: the trees of male patients in radiation group

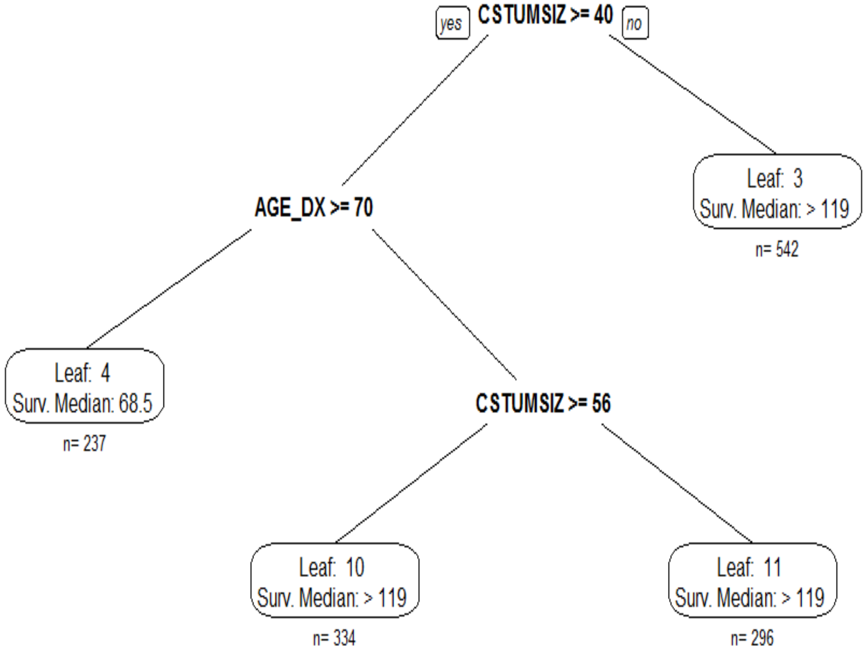


Figure 5.9: the trees of male patients in radiation with surgery group

From Figure 5.8 and 5.9, we can see that the radiation treatment group of male patients is divided into 6 subgroups from left to right, which are denoted by R1, R2, …, R6 for future analysis and both the radiation and surgery treatment group patients are divided into 4 subgroups from left to right, which are denoted by RS1, RS2, RS3, RS4. For instance, R1 group means the male patients who are aged above 76 and whose tumor size are bigger than 44 millimeters take the radiation treatment only. And R5 group indicates that the male patients who are aged below 76 and whose tumor size between 28 and 58 millimeters. Also for the patients took both radiation and surgery, the RS1 group means the male patients who are aged above 70 and tumor size are bigger than 40 millimeters. As well in Figure 5.10 and 5.11, for female patients, the radiation treatment group is divided into 7 subgroups, R1, …, R7, and the radiation and surgery group is split to 6 subgroups, RS1, …, RS6.

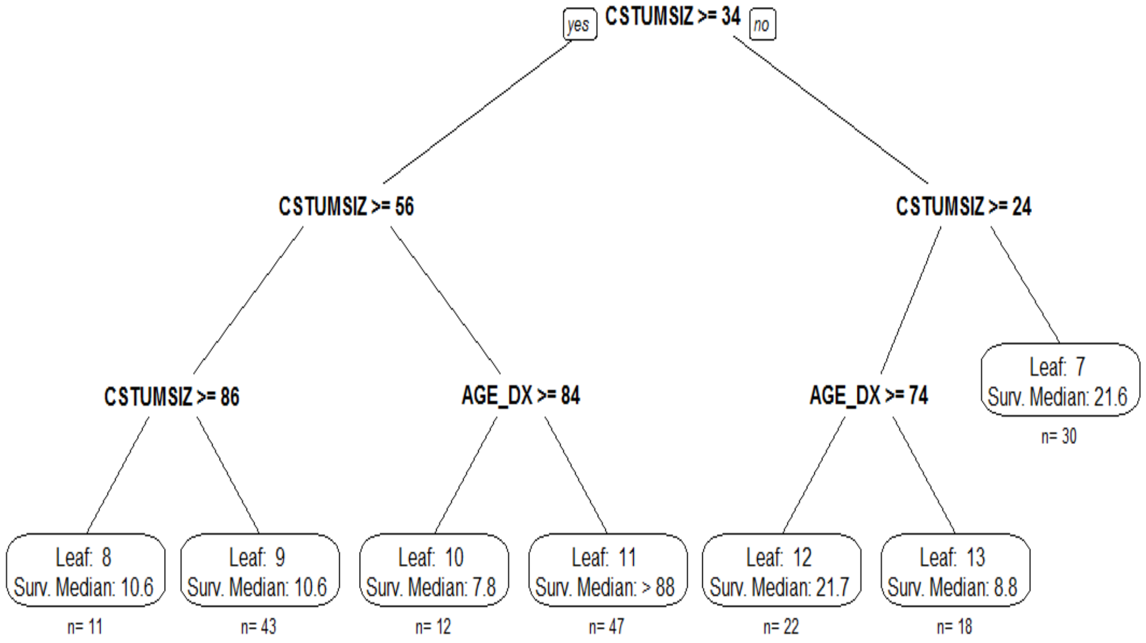


Figure 5.10: the trees of female patients in radiation group

For instance, from Figure 5.10 above, R3 group means the female patients who aged above 84 and whose tumor size between 34 and 56 millimeters take the radiation only. And R6 group means the female patients who aged less than 74 and whose tumor size between 24 and 34 millimeters.

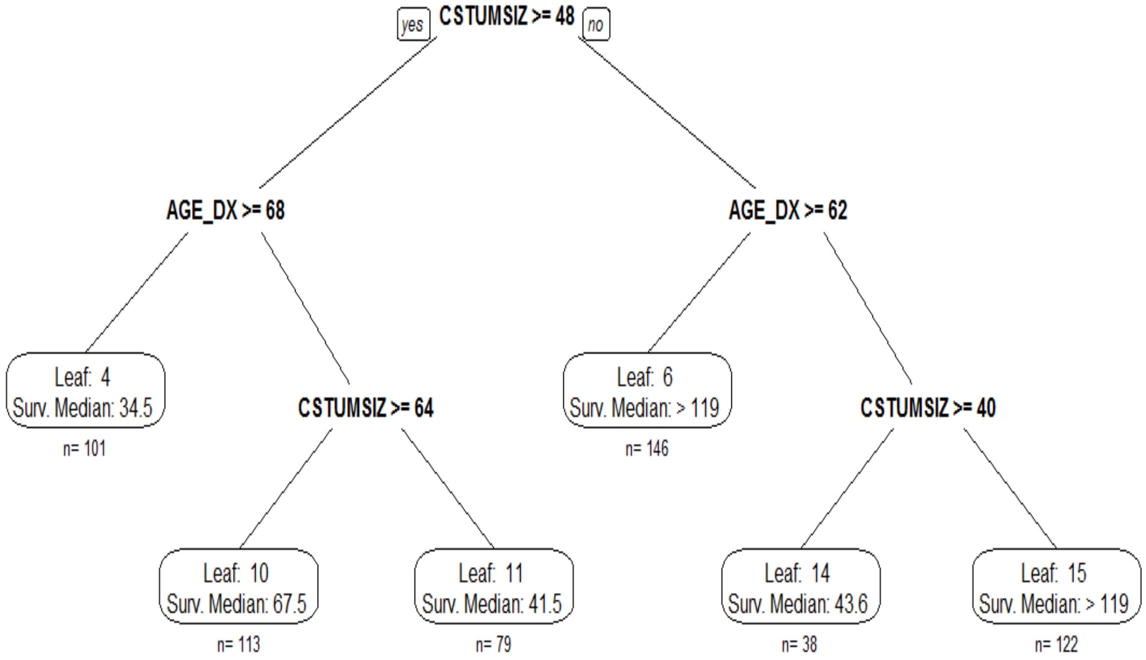


Figure 5.11: the trees of female patients in radiation with surgery group

From Figure 5.11 above, we could know that the RS1 group means the female patients who aged above 68 and with the tumor size bigger than 48 millimeters took the radiation and surgery together. And the RS5 group means the female patients who aged less than 62 with tumor size between 40 and 48 millimeters.

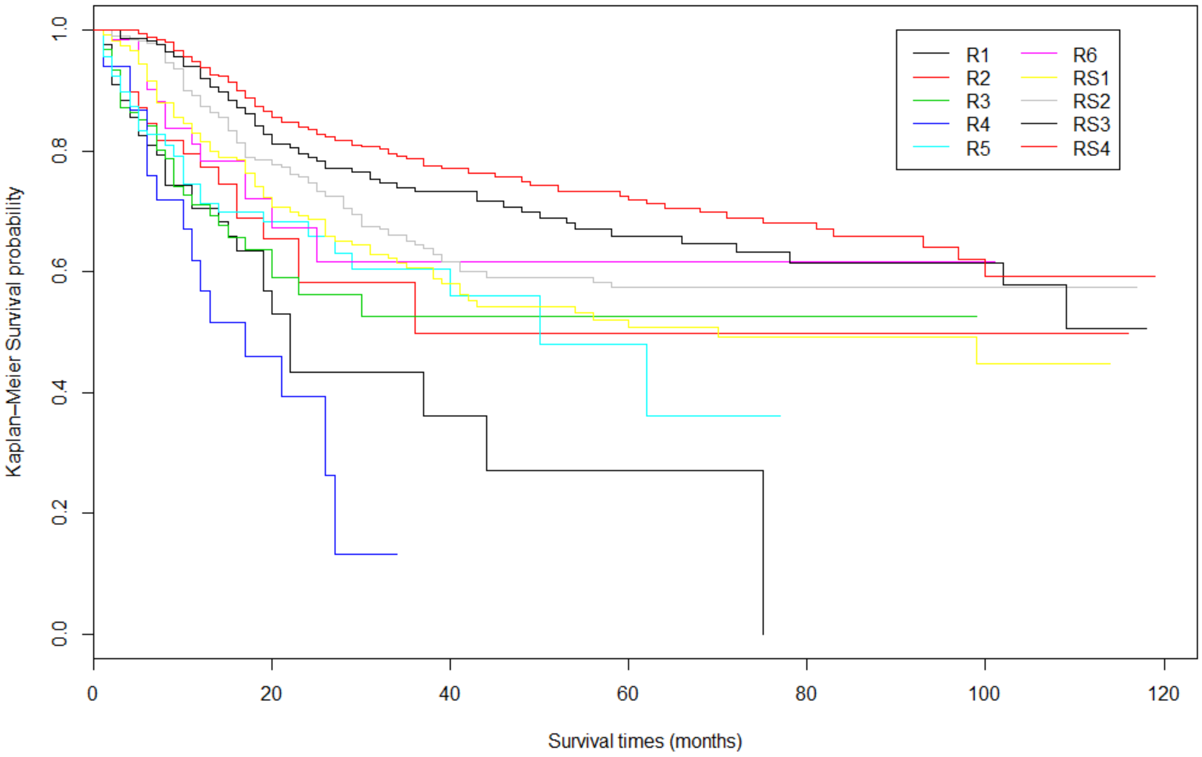


Figure 5.12: Survival curves of male patients in different subgroups

In Figure 5.12 above, we plot the Kaplan-Meier survival curves of different subgroups together to compare the treatment effect for different subgroup patients. Using the decision tree analysis, we can clear to see the survival probability for each different subgroup patients. For example, a male patient whose age is below 76 and tumor size is between 58 and 60 millimeters with radiation treatment only would have the lowest survival probability since the patients in R4 are the lowest line in Figure 5.12.

Therefore, we could group the male stomach cancer patients into three different subgroups which identify the effectiveness of treatment with radiation versus radiation and surgery by using the results of the decision tree analysis. For instance, the survival plot of RS1 is very close to R5, which suggests us that for male patients whose age is between 70 and 76 with tumor size between 40 and 58 millimeters, the combination of radiation and surgery shows no advantage over radiation only. Thus, the physicians should guide the patients not to consider receiving surgery. From Figure 5.12, we summarize below when the treatment with radiation only and radiation with surgery are almost equally effective:

(1). RS4, RS3

(2). R2, R3, R5, R6, RS1, RS2

(3). R4, R1

Thus, we found very important results which can recommend stomach doctors to give information to their patients whether they can receive radiation only instead of receiving radiation with surgery when they are equally effective to stomach cancer patients in the same age level with certain size of tumor.

Similarly, we plot the survival curves of each subgroup for female patients in Figure 5.13 by using the decision tree analysis results from Figure 5.10 and 5.11. We also group female stomach cancer patients into three clusters for identifying the effectiveness of treatment with radiation versus radiation and surgery. Below are the three clusters:

(1). RS5, RS6

(2). R6, R7, RS1, RS2, RS3, RS4

(3). R1, R2, R3, R4, R5

During each cluster, they have almost the same treatment effect for female patients, our results would help the physicians to make the decision for stomach cancer patients who could only do treatment with radiation when the patients are in the appropriate age and tumor size intervals.

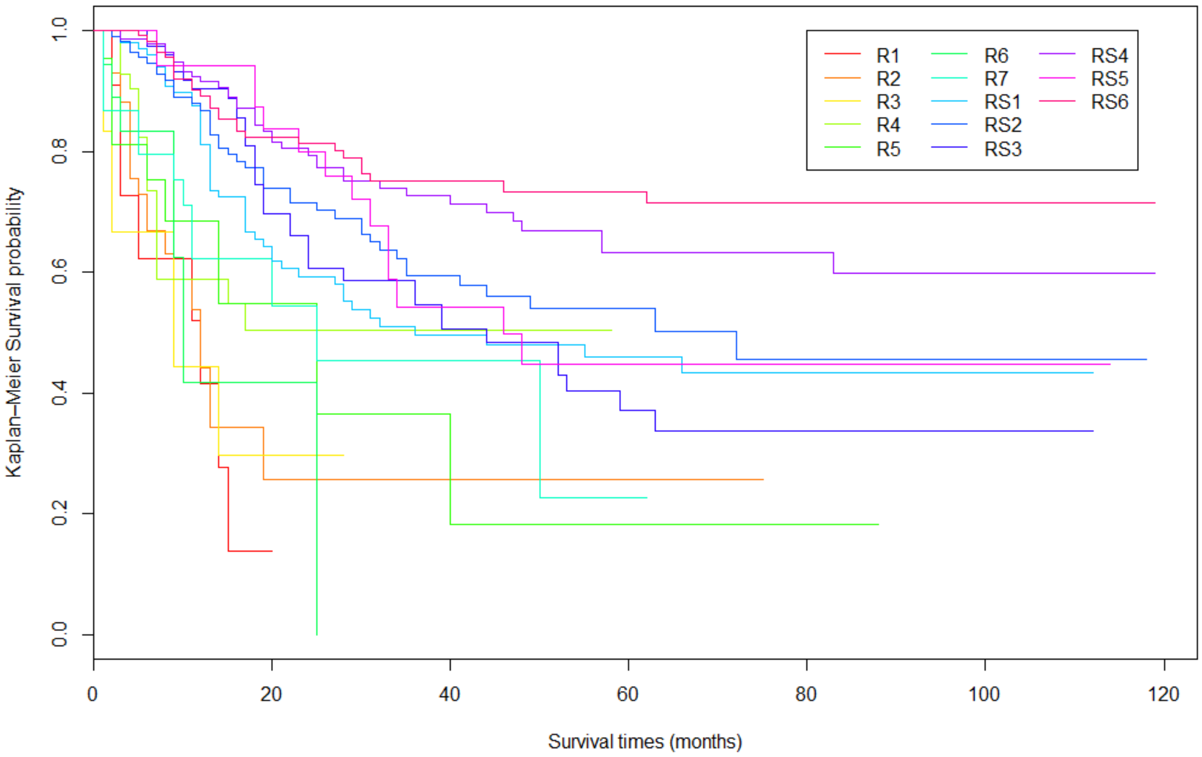


Figure 5.13: Survival curves of female patients in different subgroups

5.6 Contributions

In this study, we perform the nonparametric and parametric analysis for comparing the treatment effect for male and female stomach cancer patients, respectively. Our result shows that the patients who receive the combination of radiation and surgery have a significant effect than the patients who receive the radiation treatment alone regarding the survival time of stomach cancer patients. However, the decision tree analysis gives us the more powerful result. Based on the decision tree analysis, we found a more detailed treatment difference between the subgroups. For instance, for male patients whose age is between 70 and 76 with tumor size between 40 and 58 millimeters, the combination of radiation and surgery shows no advantage over radiation only, which can help the physicians to choose the suitable treatment for stomach cancer patients.

Chapter 6

Future Research

Stomach cancer, the third leading cause of death in the world, not much research has been done in comparing to Breast cancer, Lung cancer, Colon cancer, among others. The study in this dissertation opens up several directions for future research. One direction concerns the involvement of identifying the risk factors. What causes stomach cancer? We need to identify the risk factors that cause stomach cancer. Unfortunately, we do not have the necessary data to develop such a model to statistically identify the significant risk factors, interactions, so that we will be able to predict if a patient is a potential candidate for stomach cancer. Once such data is available we will pursuit the development of such an analytic model.

In addition to parametric analysis, we believe that Bayesian analysis is applicable in the behavior of the malignant tumor size in the stomach. Preliminary studies of the present data indicate a small significance difference in the approximate maximum likelihood estimates of the parameters that drive the probability distribution function. With the applicability of Bayesian analysis, we will improve the estimates of the malignant tumor size. For instance, we found the three-parameter Weibull probability distribution can be best describe the behavior of malignant tumor size for white male stomach patients. First, we could use bootstrapping method to select the random sample. For each sample set, we will have appropriate approximate estimates for each of the three parameters. Then, we can calculate the mean, standard deviation, variance, kurtosis, among others. Next, we can identify the one with largest variance, to be a random variable. Parametric analysis of the approximate estimates of the chosen parameter that is treated as random variable could help us to identify the prior probability distribution. For the loss function, we could use the most commonly used mean square error loss function, Higgins-Tsokos (Higgins and Tsokos, 1981) loss function. Then we proceed with Bayesian analysis to obtain Bayesian estimates of the reject parameter that will better than the parametric estimates.

Another direction concerns the modeling approach to survival analysis. Since our SEER dataset only contains patients' age, tumor size, race, gender and treatment, which is very difficult for us to apply the Cox proportional hazard model due to the strong assumptions that need to be verified. Thus, we could develop a statistical model, $R_t = f(x_1, x_2, \dots x_n)$, with the response being the time of death of a stomach cancer patient and the independent variables are the cause of him/her death. Thus, having such a statistical model, we would predict the time of death of a given patient as a function of the patients' information. Moreover, utilizing surface analysis controls the significant variables and enables us to maximize the survival time for stomach cancer patients.

References

[1] Moghimi-Dehkordi, Bijan, et al. "Statistical comparison of survival models for analysis of cancer data." Asian Pac J Cancer Prev 9.3 (2008): 417-420.

[2] Hamashima, Chisato, et al. "Survival analysis of patients with interval cancer undergoing gastric cancer screening by endoscopy." PloS one 10.5 (2015): e0126796.

[3] Moghimbeigi, Abbas, et al. "Survival Analysis of Gastric Cancer Patients with Incomplete Data." Journal of gastric cancer 14.4 (2014): 259-265.

[4] Matsuda, Tomohiro, and Kumiko Saika. "The 5-year relative survival rate of stomach cancer in the USA, Europe and Japan." Japanese journal of clinical oncology 43.11 (2013): 1157-1158.

[5] Kunz, Pamela L., et al. "Long-term survivors of gastric cancer: a California population-based study." Journal of clinical oncology 30.28 (2012): 3507-3515.

[6] Hansson, Lars-Erik, Pär Sparén, and Olof Nyrén. "Survival in stomach cancer is improving: results of a nationwide population-based Swedish study." Annals of surgery 230.2 (1999): 162.

[7] Yang, Dongyun, et al. "Survival of metastatic gastric cancer: Significance of age, sex and race/ethnicity." Journal of gastrointestinal oncology 2.2 (2011): 77.

[8] De Ville, Barry. Decision trees for business intelligence and data mining: Using SAS enterprise miner. SAS Institute, 2006.

[9] Surveillance, Epidemiology and End Results Program (2013). Cancer Statistics Factsheets: Stomach Cancer. Retrieved from http://seer.cancer.gov/statfacts/html/prost.html on August 28, 2013.

[10] American Cancer Society from http://www.cancer.org

[11] Wikipedia from http://www.wikipedia.org

[12] Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain “goodness-offit” criteria based on a stochastic processes. Annals of Mathematical Statistics, 23(2): 193-212.

[13] Chan, Y. M., Bonsu, N. O., & Tsokos, C. P. (2012). Parametric analysis of prostate cancer. Proceedings of Dynamic Systems and Applications, 6: 85-90.

[14] Chernoff, H., & Lehmann, E. L. (1954). The use of maximum likelihood estimates in chisquared tests for goodness of fit. The Annals of Mathematical Statistics, 25(3): 579-586.

[15] D. Vovoras, 2011, Statistical analysis and modeling: cancer, clinical trials, environment and epidemiology. Graduate School Theses and Dissertations. http://scholarcommons.usf.edu/etd/3397.

[16] K. Pokhrel and C. Tsokos, "Forecasting age-specific brain cancer mortality: functional
data approach," in Proceedings of Dynamic Systems and Applications, G. Ladde
and M. Sambandham, Eds., 2012, pp. 341–345.

[17] R. Koenker and B. G. Jr., "Regression Quantiles," Econometrica, vol. 46, no. 1, pp. 33–50, 1978.

[18] R. Koenker and K. F. Hallock, "Quantile Regression," Journal of Economic Perspectives, vol. 15, no. 4, pp. 143–156, 2001.

[19] M. Buchinsky, "Recent advances in quantile regression model: A practical guideline for emperical research," The Journal of Human Resources, vol. 33, no. 1, pp. 88–126, 1998.

[20] R. Koenker, Quantile Regression. Cambridge University Press, 2005.

[21] R. Koenker and V. D'Orey, "Computing regression quantile," Journal of Royal Statisticsl Society, vol. 36, pp. 383–393, 1987.

[22] ——, "A remark on computing regression quantiles," Applied Statistics, vol. 43, pp. 410–414, 1994.

[23] R. Koenker and P. Ng, "A Frisch-Newton algorithm for sparse quantile regression," Acta Mathematicae Applicatae Sinica, vol. 21, no. 2, pp. 225–236, 2005.

[24] K. Yu, Z. Lu, and J. Stander, "Quantile regression: applications and current research areas," The Statistician, vol. 52, no. 3, pp. 331–350, 2003.

[25] S. Lauridsen, "Estimation of value at risk by extreme value methods," Extremes, vol. 3, no. 2, pp. 107–144, 2000.

[26] G. Bassett and H. Chen, "Portfilio style:return-based attribution using quantile regression," Emperical Economics, vol. 26, pp. 293–305, 2001.

[27] Z. Cai, "Regression quantiles for time series ," Econometric Theory, vol. 18, pp. 169–192, 2002.

[28] A. Gannoun, J. Saracco, and K. Yu, “Nonparametric prediction by conditional median and quantiles ,” Journal of Statistical Planning Information, vol. 117, no. 2, p. 17, 2003.

[29] S. R. Lipsitz, M. G. Fitzmaurice, G. Molenberghs, and P. Zhao, “Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus ,” Journal of the Royal Statistical Society: Series C, vol. 46, no. 4, pp. 463–476, 1997.

[30] K. Yu and R. A. Moyeed, “Bayesian quantile regression ,” Statistics and Probability Letters, vol. 54, no. 4, pp. 437–447, 2001.

[31] Y.Wu and L. Liu, “Variable selection in quantile regression,” Statistica Sinica, vol. 19, pp. 801–817, 2009.

[32] R. Koenker, P. Ng, and S. Portnoy, “Quantile smoothing splines,” Biometrika, vol. 81, no. 4, pp. 673–680, 1994.

[33] Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347): 730-737.

[34] Houghton, J.C., 1978. Birth of a parent: The Wakeby distribution for modeling flood flows. Water Resources Research, 14(6), pp.1105-1109.

[35] Cohen, A.C., 1965. Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples. Technometrics, 7(4), pp.579-588.

[36] Kleiber, C., 2008. A guide to the Dagum distributions. Modeling Income Distributions and Lorenz Curves, pp.97-117.

[37] Fox, M. and Rubin, H., 1964. Admissibility of quantile estimates of a single location parameter. The Annals of Mathematical Statistics, pp.1019-1030.

[38] Barrodale, I. and Roberts, F.D., 1973. An improved algorithm for discrete l_1 linear approximation. SIAM Journal on Numerical Analysis, 10(5), pp.839-848.

[39] Karmarkar, N., 1984, December. A new polynomial-time algorithm for linear programming. In Proceedings of the sixteenth annual ACM symposium on Theory of computing (pp. 302-311). ACM.

[40] Madsen, K. and Nielsen, H.B., 1993. A Finite Smoothing Algorithm for Linear l_1 Estimation. SIAM Journal on Optimization, 3(2), pp.223-235.

[41] Cole, T.J. and Green, P.J., 1992. Smoothing reference centile curves: the LMS method and penalized likelihood. Statistics in medicine, 11(10), pp.1305-1319.

[42] Royston, P. and Altman, D.G., 1994. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Applied statistics, pp.429-467.

[43] Koenker, R. and Geling, O., 2001. Reappraising medfly longevity: a quantile regression survival analysis. Journal of the American Statistical Association, 96(454), pp.458-468.

[44] Machado, J.A. and Mata, J., 2005. Counterfactual decomposition of changes in wage distributions using quantile regression. Journal of applied Econometrics, 20(4), pp.445-465.

[45] Pokhrel, Keshav Prasad, "Statistical Analysis and Modeling of Brain Tumor Data: Histology and Regional Effects" (2013). *Graduate Theses and Dissertations.* http://scholarcommons.usf.edu/etd/4746

[46] Coad, A. and Rao, R., 2008. Innovation and firm growth in high-tech sectors: A quantile regression approach. Research policy, 37(4), pp.633-648.

[47] Hettmansperger, T.P. and Sheather, S.J., 1986. Confidence intervals based on interpolated order statistics. Statistics & Probability Letters, 4(2), pp.75-79.

[48] Nyblom, J., 1992. Note on interpolated order statistics. Statistics & probability letters, 14(2), pp.129-131.

[49] Zhou, K.Q. and Portnoy, S.L., 1996. Direct use of regression quantiles to construct confidence sets in linear models. The Annals of Statistics, 24(1), pp.287-306.

[50] Choi, Bong-Jin, "Statistical Analysis, Modeling, and Algorithms for Pharmaceutical and Cancer Systems" (2014). *Graduate Theses and Dissertations.* http://scholarcommons.usf.edu/etd/5200

[51] Bonsu, Nana Osei Mensa, "Age Dependent Analysis and Modeling of Prostate Cancer Data" (2013). *Graduate Theses and Dissertations.* http://scholarcommons.usf.edu/etd/4867

[52] Chan, Yiu Ming, "Statistical Analysis and Modeling of Prostate Cancer" (2013). *Graduate Theses and Dissertations.* http://scholarcommons.usf.edu/etd/4806

[53] Kottabi, Zahra, "Statistical Modeling and Analysis of Breast Cancer and Pancreatic Cancer" (2012). *Graduate Theses and Dissertations.* http://scholarcommons.usf.edu/etd/4350

[54] Xu, Yong, "Statistical Models for Environmental and Health Sciences" (2011). *Graduate Theses and Dissertations.* http://scholarcommons.usf.edu/etd/3414

[55] Cong, Chunling, "Statistical Analysis and Modeling of Breast Cancer and Lung Cancer" (2010). *Graduate Theses and Dissertations.* http://scholarcommons.usf.edu/etd/3563

[56] Wilcoxon, F., Katti, S.K. and Wilcox, R.A., 1963. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. American Cyanamid Company.

[57] Peto, R. and Peto, J., 1972. Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society. Series A (General), pp.185-207.

[58] Gross, A.J. and Clark, V.A., 1975. Gehan-Wilcoxon test. Survival distribution: Reliability applications in biomedical sciences, p.120.

[59] Bacchetti, P. and Segal M. R. (1995): Survival trees with timedependentcovariates: application to estimating changes in the incubation period of AIDS Lifetime Data Analysis Vol. 1, number1.

[60] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984): Classification and Regression Trees, New York; Chapman and Hall

[61] Breiman, Leo (2001): Random Forests, Machine learning, 45 (1): 5–32.

[62] Chin-Shang Li, Jeremy M.G. Taylor (2002) A semi-parametric accelerated failure time cure model. Statistics in Medicine; 21: 3235-3247.

[63] Davis, R. and Anderson, J. (1989): Exponential survival trees, Statistics in Medicine 8, pp 947-962.

[64] F. Gao, A. K. Manatunga, and S. Chen (2004), “Identification of prognostic factors with multivariate survival data”, Computational Statistics and Data Analysis 45, pp. 813-824

[65] Fabien Corbiere, Pierre Joly (2007). A SAS macro for parametric and semiparametric mixture models. Computer Methods and Programs in Biomedicine; 85(2): 173-80.

[66] Farewell, V.T. (1982) The use of mixture models for the analysis of survival data with long-term survivors. Biometrics; 38: 1041-1046.

[67] Green L. (2009) “Age Dependent Screening”, SIAM Conference Mathematics for Industry: Challenges and Frontiers, San Francisco, California.

[68] Harrington, D. P. and Fleming, T. R. (1982): A class of rank test procedures for censored survival data. Biometrika 69, 553-566.

[69] Lebalanc, M.; Crowlry, L. (1992): Relative risk trees for censored survival data, Biometrics. v48. 411-425.

[70] LeBlanc, M., Crowley, J. (1993). Survival trees by goodness of split. Journal of the American Statistical Association 88, 457–467

[71] Loh, W. Y. and Shih, Y. S. (1997): Split selection methods for classification trees. Statistica Sinica, Vol. 7, p. 815 - 840.

[72] N.A Ibrahim, et al. (2008): Decision tree for competing risks survival probability in breast cancer study, International Journal of Biomedical Sciences Volume 3 Number 1.

[73] Orbe J, Ferreira E, Nunez-Anton V. (2002) Comparing proportional hazards and accelerated failure time models for survival analysis. Statist med; 21: 3493-510.

[74] Plackett, R.L.(1983) "Karl Pearson and the Chi-Squared Test". International Statistical Review , 51 (1): 59–72.

[75] Quinlan, J. R. (1986): Induction of Decision Trees. Machine. Learning 1, 1, 81- 106.

[76] Segal M. R. (1988): Regression trees for censored data, Biometrics 44, pp.35-47.

[77] Yamaguchi K. (1992), Accelerated failure-time regression model with a regression model of surviving fraction: an application to the analysis of “permanent employment” in Japan. Journal of the American Statistical Association; 83:222-230.

[78] Rigdon, S.E. and Basu, A.P., 1989. The power law process: a model for the reliability of repairable systems. Journal of Quality Technology, 21(4), pp.251-260.

[79] Brown, M., 1972. Statistical analysis of non-homogeneous Poisson processes. Stochastic point processes, pp.67-89.

[80] Faravelli, L., 1989. Response-surface approach for reliability analysis. Journal of Engineering Mechanics, 115(12), pp.2763-2781.

[81] Higgins, J.J. and Tsokos, C.P., 1981. A quasi-Bayes estimate of the failure intensity of a reliability-growth model. IEEE Transactions on Reliability, 30(5), pp.471-475.

[82] Zhu, H.P., Xia, X., Chuan, H.Y., Adnan, A., Liu, S.F. and Du, Y.K., 2011. Application of Weibull model for survival of patients with gastric cancer. BMC gastroenterology, 11(1), p.1.

[83] Dolas, D.R., Jaybhaye, M.D. and Deshmukh, S.D., 2014. Estimation the System Reliability using Weibull Distribution. International Proceedings of Economics Development and Research, 75, p.144.