APPLICATION OF FUZZY LINEAR REGRESSION MODELS FOR PREDICTING TUMOR SIZE OF COLORECTAL CANCER IN MALAYSIA’S HOSPITAL

MUHAMMAD AMMAR BIN SHAIFI

A thesis submitted in fulfillment of the requirement for the award of the Degree of Master of Science

Faculty of Science, Technology and Human Development
University Tun Hussein Onn Malaysia

JULY 2015
ABSTRACT

Fuzzy linear regression analysis has become popular among researchers and standard model in analysing data vagueness phenomena. These models were represented by five statistical models such as multiple linear regression, fuzzy linear regression (Tanaka), fuzzy linear regression (Ni), extended fuzzy linear regression under benchmarking model (Chung) and fuzzy linear regression with symmetric parameter (Zolfaghari). A case study in colorectal cancer (CRC) data at the general hospital in Kuala Lumpur was carried out using the five models as mention above. Secondary data of 180 colorectal cancer patients who received treatment in general hospital were recorded by nurses and doctors. Twenty five independent variables with different combination of variable types were considered to find the best models to predict the size of tumor colorectal cancer. The quality of life among CRC patients which is to detect the early CRC stage is still very poor, not implemented and divulged as a nationwide programme. The main objective of this study is to determine the best model by predicting the size of tumor of CRC. Moreover, this study wants to identify the factors and symptoms that contribute the size of tumor. The comparisons among the five models were carried out to find the best model by using statistical measurements of mean square error (MSE) and root mean square error (RMSE). The results showed that the fuzzy linear regression with symmetric parameter (Zolfaghari) was found to be the best model, having the lowest MSE and RMSE value by 98.21 and 9.91. Hence, the size of tumor could be predicted by managing twenty five independent variables.
ABSTRAK

CONTENTS

TITLE i
DECLARATION ii
DEDICATION iii
ACKNOWLEDGEMENT iv
ABSTRACT v
CONTENTS vii
LIST OF TABLES x
LIST OF FIGURE xii
LIST OF ABBREVIATIONS/ SYMBOLS xiii
LIST OF APPENDICES xvii

CHAPTER 1 INTRODUCTION 1

1.1 Introduction 1
1.2 Research background of regression analysis 1
1.3 Research background of colorectal cancer 2
1.4 A brief history of fuzzy regression 5
1.5 Problem Statement 6
1.6 Research Objectives 7
1.7 The Scope of Study 8
  1.7.1 Data Scope 8
  1.7.2 Model Scope 11
1.8 Research Contribution 12
1.9 Thesis Organisation 13

CHAPTER 2 LITERATURE REVIEW 14
2.1 Introduction 14
2.2 Statistical modeling in colorectal cancer area 14
2.3 Application of fuzzy modeling 18
  2.3.1 fuzzy modeling in medical 18
  2.3.2 fuzzy linear regression models 21
2.4 Variables review 22
2.5 Summary 23

CHAPTER 3 RESEARCH METHODOLOGY 30
3.1 Introduction 30
3.2 Introduction of fuzzy sets and membership function 32
3.3 Fuzzy logic models 32
3.4 Multiple linear regression 33
3.5 Fuzzy linear regression 36
  3.5.1 Fuzzy linear regression (Tanaka, 1982) 36
  3.5.2 Fuzzy linear regression (Ni, 2005) 43
  3.5.3 Extended fuzzy linear regression under benchmarking model 43
  3.5.4 Fuzzy linear regression with symmetric parameter (Zolfaghari, 2014) 44
3.6 Cross validation technique 45

CHAPTER 4 FINDINGS AND DATA ANALYSIS 47
4.1 Introduction 47
4.2 Demographic profile of patients 47
  4.21 Categorical Variables 47
  4.22 Continuous Variables 52
4.3 Strength of the data 52
4.4 Multiple linear regression 52
  4.4.1 The variance of residuals 53
  4.4.2 The residual normally distributed 54
  4.4.3 Multicollinearity checking 55
4.5 Analysis of multiple linear regression 56
  4.5.1 Assessment for significance of individual predictor variables 56
  4.5.2 Analysis of variance (ANOVA) 58
4.6 Fuzzy Linear Regression (Tanaka) 58
4.7 Fuzzy linear regression (Ni) 61
4.8 Extended fuzzy linear regression under benchmarking models 63
4.9 Fuzzy linear regression with symmetric parameter 65
4.10 Comparison of study 68
  4.10.1 Measuring mean square error 68
  4.10.2 Comparing of root mean square error 69
4.11 Summary of results 69

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS 71
  5.1 Introduction 71
  5.2 Conclusions 71
  5.3 Recommendations for further research 74

REFERENCES 75
  Appendices A 82
  Appendices B 85
LIST OF TABLES

Table 1.1: Description of data 10
Table 2.1: Summary o literature review in colorectal cancer area 25
Table 2.2: Summary of fuzzy model in medical 28
Table 3.1: Summary of ANOVA 36
Table 3.2: Input-output data 38
Table 4.1: Socio demographic characteristics of patients 49
(n=180) for categorical
Table 4.2: Socio demographic characteristics of patients 52
(n=180) for continuous
Table 4.3: Summary stength of data 52
Table 4.4: Residual Statistic 53
Table 4.5: Coefficients of tolerance values and eigenvalues 55
and explanationvariances for actual data.
Table 4.6: Summary of parameter estimate multiple linear regression 57
Table 4.7: ANOVA for multiple linear regression 58
Table 4.8: Fuzzy parameter, Tanaka (H=0.5) 59
Table 4.9: Summary of model Tanaka 60
Table 4.10: Fuzzy parameter, Ni (H=0.5) 61
Table 4.11: Summary of model Ni 62
Table 4.12: Fuzzy parameter, Chung 64
Table 4.13: Summary of model Chung 65
Table 4.14: Fuzzy parameter, Zolfaghari (H=0.5) 66
Table 4.15: Summary of model Zolfaghari 66
Table 4.16: Result for MSE for models 68
Table 4.17: Result for RMSE for models 69
Table 5.1: MSE and RMSE value summary for all models (tomour size) 72
LIST OF FIGURES

Figure 1.1: Colon cancer and polyp 4
Figure 1.2: Stages of colorectal cancer 4
Figure 3.1: Flow chart of research framework 31
Figure 3.2: Mapping of input space to output space 33
Figure 3.3: Fuzzy set of parameter \( A : A \approx \alpha \) 38
Figure 3.4: Explanation of fuzzy linear regression model \( Y_c = A_0 + A_1x_c \) 40
Figure 3.5: Degree of fitting of \( Y_c^* \) to a given fuzzy data \( Y_e \) 41
Figure 3.6: Membership function of symmetrically triangular fuzzy number 44
Figure 4.1: Scatter plot of constant variance 53
Figure 4.2: Q-Q plot of normality 54
Figure 4.3: Normality of dependent variable 55
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Fuzzy set of A</td>
</tr>
<tr>
<td>AHP</td>
<td>Analytical hierarchy process</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>B</td>
<td>Level of existence corresponds to the level membership</td>
</tr>
<tr>
<td>BMI</td>
<td>Body mass index</td>
</tr>
<tr>
<td>CI</td>
<td>Condition index</td>
</tr>
<tr>
<td>CI</td>
<td>Confident interval</td>
</tr>
<tr>
<td>CRC</td>
<td>Colorectal cancer</td>
</tr>
<tr>
<td>c_i</td>
<td>Width of fuzzy parameter</td>
</tr>
<tr>
<td>C_{ii}</td>
<td>i\text{th diagonal element of matrix } (X^TX)^{-1}</td>
</tr>
<tr>
<td>DM</td>
<td>Data mining</td>
</tr>
<tr>
<td>EFLRBM</td>
<td>Extended fuzzy linear regression under benchmarking model</td>
</tr>
<tr>
<td>EORTC</td>
<td>European Organisation for Research and Treatment of Cancer</td>
</tr>
<tr>
<td>FCM</td>
<td>Fuzzy c-mean model</td>
</tr>
<tr>
<td>FCRM</td>
<td>Fuzzy c-regression model</td>
</tr>
<tr>
<td>FIT</td>
<td>Fecal immunochemical test</td>
</tr>
<tr>
<td>FLR</td>
<td>Fuzzy linear regression</td>
</tr>
<tr>
<td>FLRWSP</td>
<td>Fuzzy linear regression with symmetric parameter</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>FOBT</td>
<td>Fecal occult blood test</td>
</tr>
<tr>
<td>$f(x, A)$</td>
<td>Fuzzy function of set $A$</td>
</tr>
<tr>
<td>$f(X, A)$</td>
<td>Fuzzy model</td>
</tr>
<tr>
<td>gFOBT</td>
<td>Guaiac based fecal occult blood test</td>
</tr>
<tr>
<td>ICU</td>
<td>Intensive care unit</td>
</tr>
<tr>
<td>icd10</td>
<td>Place where CRC existed by patient</td>
</tr>
<tr>
<td>IDA</td>
<td>Intelligent data analysis</td>
</tr>
<tr>
<td>IRFC</td>
<td>Iterative relative fuzzy connectedness</td>
</tr>
<tr>
<td>MANOVA</td>
<td>Multivariate analysis of variance</td>
</tr>
<tr>
<td>MF</td>
<td>Membership function</td>
</tr>
<tr>
<td>MLR</td>
<td>Multiple Linear Regression</td>
</tr>
<tr>
<td>Mm</td>
<td>Milimetre</td>
</tr>
<tr>
<td>MOHM</td>
<td>Ministry of Health Malaysia</td>
</tr>
<tr>
<td>MOS</td>
<td>Malaysian Oncology Society</td>
</tr>
<tr>
<td>MRA</td>
<td>Magnetic resonance angiography</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean square error</td>
</tr>
<tr>
<td>MSR</td>
<td>Mean square regression</td>
</tr>
<tr>
<td>NCR</td>
<td>National Cancer Registry</td>
</tr>
<tr>
<td>Q-Q plot</td>
<td>Plot of quartile</td>
</tr>
<tr>
<td>R</td>
<td>Right reference function</td>
</tr>
<tr>
<td>$R^2$</td>
<td>Coefficient of determination</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean square error</td>
</tr>
<tr>
<td>RS2337107</td>
<td>Gene of polymorphism on risk colorectal cancer</td>
</tr>
<tr>
<td>Sig.</td>
<td>Significant value</td>
</tr>
<tr>
<td>SMAD7</td>
<td>Functional candidate gene for colorectal cancer</td>
</tr>
<tr>
<td>SPSS</td>
<td>Software package for statistical analysis</td>
</tr>
</tbody>
</table>
SSE - Sum of error
SSR - Sum of regression
SST - Sum of total
TNM - Tumor, nodes and metastases
VIF - Variance inflation factor
WHI - Waist circumference to height index
WHO - World Health Organization
\( \alpha \) - Center for fuzzy parameter
\( \beta_i \) - Coefficient in multiple linear regression
\( \mathbf{B}_i \) - Vector of constants
\( e \) - Index regression number from 1 to n
\( \epsilon \) - Random error of parameter
\( \mathbf{\epsilon} \) - Vector of independent normal random variables
\( \mathcal{Z} \) - Infinite dimensional feature space
\( f \) - Index regression number from 1 to n
\( H \) - Height of fuzzy triangular
\( \text{ith} \) - Sample of \( i \)
\( \mathbf{J} \) - An \( nxn \) matrix
\( n \) - Number or the respondents less than 30
\( N \) - Number of observations 30 and above
\( P \) - Amount of predictor variables
\( s \) - Symmetric measurement
\( S \) - Standard deviation
\( \mu_A(a) \) - Membership function of element \( a \) in set \( A \)
\( X_i \) - Parameter of independent variables
\( \mathbf{X}_i \) - Matrix parameters
\( \bar{x} \) - Mean or average
\( \mathbf{X}^\intercal \) - Transpose for matrix \( \mathbf{X} \)

\( Y \) - Dependent variables/ observations

\( y_i \) - Observation of data

\( \hat{y} \) - Output of fuzzy model

\( \hat{y}_i \) - Predicted data

\( y^*_{i} \) - Estimated value for \( y \)

\( \mathbf{Y}_i \) - Vector of responses

\( \Sigma \) - Covariance matrix
LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>List of publications</td>
<td>81</td>
</tr>
<tr>
<td>B</td>
<td>Data for CRC patients in general</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Hospital around Kuala Lumpur</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Introduction

This chapter presents the introduction on research background of regression analysis, colorectal cancer, history of fuzzy regression, problem statement, research objectives, scope of study, research contribution and research organization. Lastly, a brief summary is included at the end of chapter.

1.2 Research background of regression analysis

Regression analysis has become one of the standard models in analysing data. It’s popularity comes from several sources. The statistical equation is derived obtained from the analysis which explains the relationship of dependent and independent variables. It provides much explanatory power, especially due to its multivariate nature. It is widely available in computer packages and can be easily interpreted. It is extensively used in applied sciences, economic, engineering, computer, social sciences and other fields (Agresti, 1996).

Nonlinear modeling is of interest to many researchers in modeling, statistics, rather than linear modeling. The obtained functional form should be approximately close to the real data. If the functional form varies much from the real data, the estimation will be inconsistent, bias and so on. However, other difficulties may also arise with non-linearity approach (Rousseeuw et al., 1987). Nowadays, there are many modeling resulting from the regression analysis, such as multiple regression, quadratic regression, cubic regression, logit model, probit model, exponential model, growth model, neural network regression and fuzzy regression.
1.3 Research background of colorectal cancer

Colorectal cancer (CRC) is a term of cancer in the colon and/or rectum. The colon and rectum are two different parts in the human body which play an important role to digest food and past waste. Colorectal cancer is one of the commonest diseases malignancies in the world (Malaysian Oncological Society, 2007). According to World Health Organization (WHO) (2012), colorectal cancer was the fourth leading cause of death among cancer patients. The risk of this particular cancer is rising in many countries, especially in the Asian Region of Malaysia (Yusoff et al., 2012).

Colorectal cancer is the third leading cause of cancer deaths in Malaysia. According to the data of 2010 from WHO, lung cancer is the highest rated cause of cancer deaths by 17.93% and followed by breast cancer 15.83% and colorectal cancers 13.10%. Data from the Ministry of Health of Malaysia showed that in 1995, colorectal cancer admission rates increased from 8.1% in 1987 to 11.9%. Recent studies also showed an incremental trend in colorectal cancer among Asian population (Wendy, 2008 and Radzi, 2008). According to Second Paper of the National Cancer Registry (2003), there were 14.2% of male and 10.1% of female colorectal cancer patients were identified in Malaysia. It can be concluded that colorectal cancer is the commonest cancer among men and the second most common cancer among women.

At present, cause of CRC is not completely clear. However, it involves many factors contributing in developing the cancer such as increasing age, lack of nutrition, family history, body weight and so on. People with colorectal cancer may develop a number of non-specific symptoms. Symptoms of colorectal cancer typically include rectal bleeding and anemia which are sometimes associated with weight loss and changes in bowel habits (Cappell, 2005).

Colorectal cancer can be detected by factors of symptoms or common symptoms, namely abdominal cramps, a change in your bowel habits, blood in stools, weight loss, fatigue and loss of appetite. All these symptoms may show their activation in one month after the colorectal cancer eventuated (Vanagunas, 2011). According to American Cancer Society (2014), the cancer can be examined by physical exam test to check for early symptoms. The test includes colonoscopy, fecal
occult blood testing, sigmoidoscopy, biopsy and complete blood count. CRC includes four stages from an early stage until final stage. Stage I indicates that the cancer is confined to the inner lining of the colon or rectum. Stage II is cancer spreads through the wall of the colon or rectum while in stage III, cancer spreads to nearby lymph nodes and the final stage, IV is cancer spreads to distant parts of the body, such as liver or lungs (Malaysian Oncological Society, 2007).

European Organisation for Research and Treatment of Cancer (EORTC) in 2014 stated that the colorectal cancer (CRC) is also increasing significantly in other countries. It is the second most common cancer in the United Kingdom after lung cancer as well as in the United States of America. More than 50,000 people from each year dies of colorectal cancer (Jemal et al., 2008). Colorectal cancer is the leading killer among smokers in the world (Frieden et al., 2002). In the Asian population, there is a rapid increase in the trend of CRC, where it becomes the most common cancer among male superseded lung cancer (National Cancer Registry, 2006).

In Asia, screening with faecal occult blood test is a national policy only in Japan, Taiwan and Korea (Hyodo et al., 2010). Currently, in Malaysia, the awareness of the importance of CRC screening is very low, especially among the general population as well as the policy makers resulting in inadequate resources allocation for faecal occult blood test and colonoscopy. The advances in the treatment of colorectal cancer and effectiveness of the screening program, especially in the developed countries resulted increasing number of colorectal cancer survivors. Not only the disease could be cured but patients can also survive longer with the disease. Therefore, the paradigm of outcome in colorectal cancer treatment in the past decade has slowly shifted towards improving the patients’ quality of lives, survival and disease free survival. This shows that health is related to quality of life and it is one of the fundamental aspects in colorectal cancer management while survival and disease free survival remains (Bottomley, 2002).

In Malaysia, about 1 in 15 people develop colon cancer and the mean age for colorectal cancer is 61 years (range 15 to 95 years) and the majority patients were 50 years and above (National Cancer Patient Registry, 2010). Patients with colorectal cancer have different pathways of presentation for diagnosis; primary screening, symptomatic and incidental. Most patients presented symptomatically. The most common presenting symptom is “diarrhea, constipation, or other change in bowel
habit”, followed by “weight loss” and “abdominal pain”. The rectum was the most frequent site of the primary tumor followed by sigmoid colon and the rectosigmoid. Figure 1.1 and Figure 1.2 shows the tumor sites and stages.

Figure 1.1: Colon cancer and polyp

Figure 1.2: Stages of colorectal cancer
1.4  Research background on fuzzy regression

A fuzzy regression model is used in evaluating the functional relationship between the dependent and independent variables in a fuzzy environment. In many cases of fuzzy regression, the linear regression is recommended for practical situations when decisions often have to be made on the basis of imprecise and/or partially available data. Several methods have been presented to estimate fuzzy regression models. At first, fuzzy regression was developed by Tanaka et al. in year 1982 for linear case by focusing on extension principle (Taheri, 2003).

Making sense of data is an ongoing task for researchers and professionals in almost every practical endeavor. The age of information technology, characterized by a vast array of data, has enormously amplified this quest and made it even more challenging. Data collection has become the reality of our lives at any time and from everywhere. It is reported that understanding the data, revealing underlying phenomena, and visualizing major tendencies are the major undertakings to pursue in intelligent data analysis (IDA), data mining (DM), and system modeling (Pedryoz, 2005).

Fuzzy regression used in complex systems such as in industry, economy, finance, marketing, and ecology function in the real world and it is more imprecision. Such systems require decisions based on human thinking and judgmental and involve human–machine interactions. In such environments, human often not able to obtain exact numerical data about the system. The nature of information about the complex systems with vagueness is frequently fuzzy. In general, fuzzy regression seems to be intuitively more adequate for real life problems. Therefore, fuzzy regression analysis is more effective for modeling of complex systems. The pioneering work in this field reported that the authors used Zadeh’s extension principle, α-level procedure, interval arithmetic, and linear programming techniques to develop a fuzzy linear regression analysis. Minimization of these distances in the fuzzy number space with respect to the unknown parameters of regression models leads to solving systems of equations (Aliev et al., 2002).
1.5 Problem statement

This research is carried out to detect size of the tumor of colorectal cancer which are in stage I and stage II in Malaysia. There are several studies about colorectal cancer (Varela, 2010). However, solutions for early stage detection is still not found.

In Malaysia, only a few studies were done to access the quality of life among CRC patients. Natrah (2012) conducted a study to determine the quality of life in CRC patients, especially those under treatment. The respondents in this study were mainly from stage III and IV, which are considered the tumor size in final stages of the disease and no respondents were from stage I and II from the public tertiary level general hospitals. This finding shows that the awareness of tumor size of colorectal cancer screening in Malaysia to detect the early stage of CRC is still very poor and not implemented as a nation’s wide program. This correlates with findings from a study made by Kong et al. (2010), where it claimed that it is close to zero awareness of CRC screening tumor size among general population.

People, whose family members have colorectal cancer, especially at an old age, have a higher risk of developing this disease. A woman who has had cancer of the ovary, uterus or breast is more likely to develop colorectal cancer. Many people do not know about the CRC program and also there are people who know that they are having the symptoms of CRC but refuse to go for screening. For the screening program to be effective, an “at-risk” population has to be identified. The incidence of colorectal cancer increases exponentially with age and current figures from the developed world show that 90% of cases arise from those aged over 50 years old. In developed nations, the condition is thus too rare below the age of 50 years old to make screening test (Bennet, 1996).

In fuzzy linear regression, there are important models such as fuzzy regression and multiple linear regressions. Both models are usually used by researchers with continuous dependent data. Fuzzy linear regression models were focused to get the factors and symptoms that give influence to tumor size and to get the less error of model. Recently, researchers are interested to get small value of error in analysis by using fuzzy models (Saifullah et al., 2012). Hence, fuzzy linear regression is the effective model to adopt in this study based on previous research.
1.6 Research objectives

This research is an attempt to find the most appropriate colorectal cancer model in order to predict the size of tumor among CRC patients. The specific objectives of this study are detailed as below:

(i) To apply the existing latest model of fuzzy linear regression with symmetric parameter in order to identify the factors and symptoms of CRC patients.

(ii) To apply the existing models which are multiple linear regression, fuzzy linear regression (Tanaka), fuzzy linear regression (Ni), and extended fuzzy linear regression by benchmarking models under fuzziness (Chung) in order to identify the factors and symptoms of CRC patients.

(iii) To make comparison among the linear regression models as mention above in order to find the best model for predicting tumor size in colorectal cancer.

(iv) To identify the factors and symptoms that give most significant influence to the colorectal cancer.
1.7 The scope of the study

The scope of the study will discuss about two sections which are the scope of the data and followed by the scope for the model.

1.7.1 Data scope

This study used secondary data obtained from the general hospital in Kuala Lumpur. This data consists of 180 patients as respondents for colon cancer. It was recorded by nurses and doctors using cluster sampling. Dependent variable or outcome is the size of tumor. Actually, there were twenty five independent variables which are gender, ethnic, age, classification of diseases and related health problems (icd10), tumor nodes metastases (TNM) staging, family history of colon cancer, diabetes mellitus, crohn's disease, ulcerative colitis, polyp, history of cancer, endometrial, gastric, small bowel, hepatobiliary, urinary tract, ovarian, other cancer, intestinal obstruction, colorectal, weight loss, diarrhea, anaemia, blood stool and abdominal pain.

The range value size of the investigated tumor is 20mm till 100mm. The patient’s feedback on colorectal cancer was interviewed face to face by doctors and the answers were obtained immediately. The questions covered twenty four variables except for TNM staging.

Stage of colorectal cancer focuses on how big the cancer is and whether it is spread in the bowel of human body from polyp. Doctors and nurses used test and scanned to determine the stage of colon cancer faced by the patients. TNM stands for tumor, node and metastases. This way is widely used by medical cancer fields in the world. TNM describes the size of a primary tumor (T), any lymph contain cancer cells (N) and the spread of cancer to another part of body (M).

There are four stages in a primary tumor (T) which classified as, T1: tumor is only in the inner layer of the bowel, T2: the tumor is grown into the muscle layer of the bowel wall, T3: the tumor is grown into the outer lining of the bowel wall, T4: the tumor is grown through the outer lining of the bowel wall. While, three possible stages for lymph contain cancer cells identified in the lymph nodes (N). They are categorized as N0: there are no lymph nodes containing cancer cells, N1: 1 to 3 lymph nodes close to the bowel contains cancer cells, N2: there are cancer cells in 4 or more nearby lymph nodes. Furthermore, there are two stages of cancer spread (M)
which are M0 indicating that the cancer is not spread to other organs, M1 denotes that the cancer is spread to other parts of the body (American Cancer Society, 2012).

Doctors and nurses used the number systems of TNM stages of colorectal cancer patients to classify them into four main stages. Stage I defines the cancer has grown through the inner lining of the bowel or into the muscle wall. There is no cancer in the lymph nodes (T1, N0, M0 or T2, N0, M0). Stage II is divided into two categories as IIA and IIB. Stage IIA is the cancer has growing into the outer covering the bowel wall but there are no cancer cells in the lymph nodes (T3, N0, M0). Contradict, stage IIB is the cancer has growing through the outer covering of the bowel wall and into tissues or organs next to the bowel. But no lymph nodes are affected and the cancer does not spread to other area of the body (T4, NO, M0). Furthermore, stage III is divided into three stages comprising IIIA, IIIB and IIIC. Stage IIIA is the cancer is still in the inner layer of the bowel wall or has growing into the muscle layer and between 1 and 3 nearby lymph nodes containing cancer cells (T1, N1, M0 or T2, N1, M0), Stage IIIB is the cancer has growing into the outer lining of the bowel wall or into surrounding body tissues or organs, and between 1 and 3 close to lymph nodes containing cancer cells (T3, N1, M0 or T4, N1, M0).Stage IIIC defines the cancer can be of any size and is spread to 4 or more nearby lymph nodes, but there is no cancer spread to any other part of the body (any T, N2, M0). Lastly, stage IV is the cancer has spread to other parts of the body such as the liver or lungs through the lymphatic system or bloodstream (any T, any N, M1).
Table 1.1: Description of data

<table>
<thead>
<tr>
<th>No</th>
<th>Variable Name</th>
<th>Variable Type</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tumor size</td>
<td>Quantitative discrete with Minimum size= 20mm and Maximum size= 100mm</td>
<td>Size of tumor by patient</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>Qualitative binary with Female and Male</td>
<td>Gender of patient</td>
</tr>
<tr>
<td>3</td>
<td>Age</td>
<td>Quantitative discrete with Minimum age= 21 and Maximum age= 90</td>
<td>Age of patient</td>
</tr>
<tr>
<td>4</td>
<td>Ethnic</td>
<td>Qualitative categories with Malay, Chinese and Others</td>
<td>Race of patient</td>
</tr>
<tr>
<td>5</td>
<td>Family history</td>
<td>Qualitative binary with False and True</td>
<td>Family history of patient</td>
</tr>
<tr>
<td>6</td>
<td>Small bowel</td>
<td>Qualitative binary with False and True</td>
<td>Size small bowel of patient</td>
</tr>
<tr>
<td>7</td>
<td>Weight loss</td>
<td>Qualitative binary with False and True</td>
<td>Loss weight of patient</td>
</tr>
<tr>
<td>8</td>
<td>Diarrhea</td>
<td>Qualitative binary with False and True</td>
<td>Three or more loose or liquid stools per day by patient</td>
</tr>
<tr>
<td>9</td>
<td>Anemia</td>
<td>Qualitative binary with False and True</td>
<td>Decrease the red blood by patient</td>
</tr>
<tr>
<td>10</td>
<td>Blood stool</td>
<td>Qualitative binary with False and True</td>
<td>Blood in the faeces</td>
</tr>
<tr>
<td>11</td>
<td>Abdominal pain</td>
<td>Qualitative binary with False and True</td>
<td>Symptom faced by patient</td>
</tr>
<tr>
<td>12</td>
<td>icd10</td>
<td>Qualitative multinomial with Rectum Colon and Restosigmoid junction</td>
<td>Place where CRC existed by patient</td>
</tr>
<tr>
<td>13</td>
<td>TNM</td>
<td>Qualitative stages of cancer with I, IIA, IIB, IIIA, IIIB, IIIC, IV</td>
<td>Stages of CRC faced by patient</td>
</tr>
<tr>
<td>14</td>
<td>Diabetes Mellitus</td>
<td>Qualitative binary with False and True</td>
<td>Diabetes of patient</td>
</tr>
<tr>
<td>15</td>
<td>Crohn's Disease</td>
<td>Qualitative binary with False and True</td>
<td>As syndrome Crohn</td>
</tr>
<tr>
<td>16</td>
<td>Ulcerative collitis</td>
<td>Qualitative binary with False and True</td>
<td>Colitis ulcerosa or bigger collitis</td>
</tr>
<tr>
<td>17</td>
<td>Polyp</td>
<td>Qualitative binary with False and True</td>
<td>Abnormal growth of tissue</td>
</tr>
<tr>
<td>18</td>
<td>History of cancer(s)</td>
<td>Qualitative binary with False and True</td>
<td>Patient who ever in CRC or not before</td>
</tr>
<tr>
<td>19</td>
<td>Endometrial</td>
<td>Qualitative binary with False and True</td>
<td>Cancer that arises from the endometrium</td>
</tr>
<tr>
<td>20</td>
<td>Gastric</td>
<td>Qualitative binary with False and True</td>
<td>Digestive fluid that formed in the stomach</td>
</tr>
<tr>
<td>21</td>
<td>Urinary tract</td>
<td>Qualitative binary with False and True</td>
<td>Acute cystitis or bladder infection</td>
</tr>
</tbody>
</table>
Table 1.1 (Continued): Description of data

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Data Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Hepatobiliary</td>
<td>Qualitative binary</td>
<td>Gall bladder and bile ducts and do a bile</td>
</tr>
<tr>
<td>23</td>
<td>Ovarian</td>
<td>Qualitative binary</td>
<td>Cancer develops from the ovary</td>
</tr>
<tr>
<td>24</td>
<td>Other cancer</td>
<td>Qualitative binary</td>
<td>Other cancer faced by patients</td>
</tr>
<tr>
<td>25</td>
<td>Intestinal Obstruction</td>
<td>Qualitative binary</td>
<td>Preventing the normal transit of the products of digestion</td>
</tr>
<tr>
<td>26</td>
<td>Colorectal</td>
<td>Qualitative binary</td>
<td>Development of cancer in the colon or rectum</td>
</tr>
</tbody>
</table>

1.7.2 Model Scope

Firstly, data filtering should be carried out to determine either the data is suitable or not suitable to use in colorectal cancer case. In this study, we will use the coefficient of determination, R² and R-value to check the correlation in a data set that is accounted for by a statistical model.

Then, multiple linear regression should fulfill the assumptions before proceeding with the analysis. The assumptions are made as constant variance, normality distributed and multicollinearity. Thereafter, analysis will be applied by finding the significant variables, mean square error and root mean square error. The significant variables are important to identify the factors and symptoms to provide a higher effect in colorectal cancer prediction.

Moreover, other models also will be applied such as fuzzy linear regression (Tanaka, 1982), fuzzy linear regression (Ni, 2005), extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012) and fuzzy linear regression with symmetric parameter (Zolfaghari, 2014) models. All these models involve variables gender, age, ethnic, icd10, TNM staging, family history of colon cancer, diabetes mellitus, crohn's disease, ulcerative colitis, polyp, history of cancer, endometrial, gastric, small bowel, hepatobiliary, urinary tract, ovarian, other cancer,
intestinal obstruction, colorectal, weight loss, diarrhea, anemia, blood stool and abdominal pain.

The comparison among multiple linear regression and four models of fuzzy linear regression will be based on the mean square error (MSE) and root mean square error (RMSE). The model with the smallest value of MSE and RMSE will be chosen as the best model. In order to get the best model, it is important to obtain the approximation of the solution which is closed to predict the size of tumors, especially in the early stage (stage I and II) faced by patients.

1.8 Research contribution

The benefits of this study can contribute to the general hospital oncology specialist. The contributions are stated as follows:

(i) The application of fuzzy linear regression using continuous data produces accurate prediction.
(ii) The suggestion of mean square error and root mean square error using the best model could be as high prediction tumor size of CRC patient. Hence, medical decision in Oncology Department about predicting tumor size of CRC can be more reliable and accurate, especially in early stage of CRC.
(iii) The application of multiple linear regression, fuzzy linear regression (Tanaka 1982), Fuzzy Linear Regression (Ni, 2005), extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012) and fuzzy linear regression with symmetric parameter (Zolfaghari, 2014) for the colorectal cancer in the general hospital.
(iv) The recommendation based on the best model of regression in achieving better prediction about the tumor size of colorectal cancer by factors and symptoms in general hospital. The best model will be applied in Malaysia. Hence, it will change the ranking of danger in the group of cancer. Moreover, it will help Oncology Department to reduce cost for treatment and instruments for medical cases.
1.9 Thesis organization

This study comprises of five chapters. Chapter 1 is the introduction of the thesis. This chapter contains research background of regression analysis and colorectal cancer, history of fuzzy regression, problem statement, research objectives, scope of the study, research contribution and research organization.

Chapter 2 is the literature review that reviews the current and previous researches on medical field, especially on colorectal cancer in the general hospital. The applications of fuzzy linear regression also presented in medical field.

Chapter 3 is a detailed explanation about five models used in this study. The discussed models are multiple linear regression, fuzzy linear regression (Tanaka 1982), fuzzy linear regression (Ni, 2005), extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012) and fuzzy linear regression with symmetric parameter (Zolfaghari, 2014).

Chapter 4 discusses the analysis of the applied five models. The results of significant variables from multiple linear regression, mean square error and root mean square error for five models is carried out. The procedure of the analysis must be fulfilled to get the best model to predict the size of tumor by factors and symptoms variables.

Lastly, Chapter 5 concludes and summarizes the study of the prediction of colorectal cancer in general hospital and discusses the results and findings. The best model which has the lowest MSE and RMSE value is carried out and that model is selected for colorectal cancer. Recommendations to improve prediction for future research in colorectal cancer cases are also included in this chapter.
CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the review of several relevant studies on factors and symptoms that affected patients of colorectal cancer from the general hospital in Kuala Lumpur. This chapter also presents the review of the literature on colorectal cancer in data, included review of statistical analysis on colorectal cancer study and review application of fuzzy analysis. Colorectal cancer is a common disease happens in Malaysia and other countries. Usually, researcher uses basic statistics to analyze case of colorectal cancer such as logistic regression, descriptive analysis or linear regression and statistics modeling. Since the basic statistics analysis usually used in the research, the improvement statistic in analysis of colorectal cancer will be applied such as fuzzy linear regression as the latest model of linear regression in statistic.

2.2 Statistical modeling in colorectal cancer area

Brand et al. (2006) studied the risk of colon cancer in hereditary non-polyposis CRC and predicted by fuzzy modeling. The aim of the study is to identify if fuzzy logic model can be used to predict CRC risk especially in smokers. Fuzzy logic method is used for analysis. The general results showed that smoking habit significantly increases the CRC risk, usually among male. They also said fuzzy modeling may enable formulation of clinical risk scores and allows individualization of CRC prevention strategies.

Varela et al. (2010) studied to understand factors related to CRC screening among urban Hispanics and focus groups were conducted. This study has been analyzed CRC among Hispanics using cross-sectional study and descriptive analysis.
The main goal of the study was to understand the factors contributing to CRC and the results stated that the facilitators and barrier are the main factors toward CRC.

Stefani et al. (2011) researched on dietary patterns and risk of colorectal cancer in Uruguay. They examined the risk of colorectal cancer, such as prudent, traditional, drinker and others using logistic regression and polynomial regression. The factors, prudent, western and drinker are the main factors for colorectal cancer to eventuate.

Pereira et al. (2011) explored anxiety, depression, traumatic stress and quality of life in colorectal cancer after different treatments. The purpose of the study is to examine the impact of treatment on anxiety, depression, traumatic stress and quality of life in colorectal cancer patients. 114 samples of oncology patients from Hospital in the North of Portugal were used. The samples used three modes of treatment: surgery, surgery plus chemotherapy or surgery followed by radiotherapy. Measurements of the study included gender, anxiety, depression, traumatic stress and quality of life by using MANOVA test. The findings exposed that the patients who received only surgery as treatment had lower levels of depression, anxiety and traumatic stress symptoms compared to patients who received surgery and chemotherapy or surgery plus radiotherapy. They also discussed that the research of colorectal cancer illness is important and need to be improve over a year to get the lower risk of cancer.

Qureshi et al. (2012) evaluated the effective mass screening for colorectal cancer in Malaysia based on consensus or clinical practice guidelines. They recorded the race and region as variables and used descriptive analysis as methodology. The results of the study concluded that low incidence of colorectal cancer in Malaysia for mass screening is still low for average risk individuals.

Yusoff et al. (2012) studied participation and barriers to colorectal cancer screening in Malaysia. They stated colorectal cancer is the most common cancer in males and the third most common in females. The aim of this study is to detect the colorectal cancer screening and barriers among individuals in Malaysia by using descriptive analysis. There are six variables used which are age, race, education, occupation, self-reported medical illness and smoking status. The results showed that the participation in colorectal cancer screening in Malaysia is extremely low and multiple factors contribute to colorectal cancer. They also recommended that efforts should be made to increase the screening for colorectal cancer in Malaysia.
Roslani et al. (2012) researched on screening for colorectal neoplasias with fecal occult blood test: false positive impact of non-dietary restriction. The main objective of the study was to evaluate and compare the validity and suitability between two screening methods, which are Fecal Immunochemical Test (FIT) and Guaiac Based Fecal Occult Blood Test (gFOBT) rely on detection of fecal occult blood (FOBT). gFOBT is common method that used in colorectal cancer screening and detected any pseudo peroxidase activity in the feces and this method is cheaper and suitable in Malaysia for screening rather than FIT. It requires large number of patients of colorectal cancer in Malaysia. The use of FIT is to detect intact globin protein portion of human hemoglobin and more specific screening test but it is costly. They use only descriptive analysis for demographic part and use several variables such as age, race, sex, types of fecal occult blood and others. Furthermore, the result showed that FIT is the best screening with low-false of positivity and better sensitivity and specific rates.

Lavdaniti et al. (2012) developed an evaluation of depression level in colorectal cancer patients. They stated colon cancer is the third most common cancer in male and female and depression is the most factor related to colorectal cancer. They used sample in Greece with lowest number of cases of colon cancer in Europe. The aim of the study was to get the frequency of depression among patient in colon cancer and to evaluate the contributing factors. Depression and comorbid disabling syndrome are the most influential factors with 15-25% of effects on colorectal cancer. The variables that they used were gender, age, profession, educational level, marital status, place of residence, time since diagnosis and score of depression by using descriptive analysis and multiple linear regression with stepwise method. After analysis done, they got score for depression as dependent variable and the results showed it was no significant variables related to score of depression. It is concluded that depression has no effect on colorectal cancer. They recommended for more researches to be carried out in the future to determine the patterns of depression and characteristics in colorectal cancer.

Obi and Imianvan (2012) proposed fuzzy neural approach for colon cancer prediction. The proposed system is self-learning and adaptive and able to handle the uncertainties which often associated with the diagnosis and analyse the colon cancer. There are nine symptoms need to be diagnosed before designing the fuzzy neural which includes severe abdominal pains, high fever (>101 degrees Fahrenheit),
shortness of breath, red or dark blood in stool, constipation, cramps or bloating, narrow stool, weight loss, and change in level of alertness. They also recommended to create fuzzy logic membership functions that defines the value of input/output terms used in the rules.

Phillips et al. (2013) studied on initiating the colorectal cancer screening at the age of 50. Their aim was to get the factors associated with colorectal cancer screening among persons age 50-75 by using descriptive and logistic regression analysis. Independent between categorical variables, chi-squared and one-way ANOVA methods were applied in this study. The variables included in this study were screening initiation, health status indicators, healthcare utilization, age, sex, education and race. The results showed almost 75% of the 50 years old and above had initiated colorectal cancer. The researchers in this study suggested that health care in colorectal cancer screening must be encouraged according to guidelines.

Nishihara et al. (2013) studied on long-term colorectal-cancer incidence and mortality after lower endoscopy. The purpose of the study was to examine the lower endoscopy with colorectal cancer incidence and colorectal cancer mortality among patients. The variables used were gender, sex, baseline history of cancer, ulcerative colitis, colorectal polyps, familial polyposis syndromes and previous lower endoscopy by using descriptive analysis and cox-proportional hazards model (multivariate model) to evaluate hazard ratio and 95% confidence intervals. The general findings showed that the colonoscopy and sigmoidoscopy were associated to reduce the incidence of colorectal cancer and only colonoscopy associated to reduce the mortality colorectal cancer among patients.

Akbari et al. (2014) searched the lack of influence of the SMAD7 gene RS2337107 polymorphism on risk of colorectal cancer among Iranian population. SMAD7 is identified as functional candidate gene for colorectal cancer and RS2337107 as gene of polymorphism on risk CRC. The aim of the study was to investigate the influence of SMAD7 gene and RS2337107 on the risk of CRC towards clinic pathological features in Iranian population. On the other hand, this study aimed to analyze statistically the relationship of risk of colorectal cancer and clinic pathological properties. This study used 105 patients of colorectal cancer and logistic regression as their model. The variables that they used are sex, age, smoking status, grade of colorectal cancer, TNM staging, and duke’s stage. The results revealed that there are no significant associations between SMAD7 gene and
RS2337107 on the risk of CRC with clinic pathological features. They concluded that SMAD7 gene and RS2337107 are not potential contributors to the risk of colorectal cancer and suggested for a large scale study of colorectal cancer to validate their results.

Kaneko et al. (2014) studied on new index of abdominal obesity which effectively predicts the risk of colon tumor development among female Japanese. Obesity caused by accessing intake of fat-rich foods and less exercise which is known as statistically significant risk factor for colorectal cancer. The aim of the study is to obtain relationship between abdominal obesity and colorectal cancer tumor among CRC patients. 657 of patients were used as samples of the study. Several variables used such as age, sex, waist circumference to height index (WHI score), body mass index (BMI score) and others. Descriptive analysis and logistic regression were chosen in the study. From the results, the general conclusion of the study can be concluded as the risk factors for colon tumor are different between women and men and WHI is a main factor for women to get colon tumor among Japanese.

2.3 Application of fuzzy modeling

Many researchers have used fuzzy modeling in many fields to analyze cases such as in medical, science, engineering and others. Commonly, fuzzy modeling used to analyze the cases with more complexes and it is reliable. Several previous studies used fuzzy modeling since fuzzy created by Lotfi A. Zadeh in 1965 with fuzzy set theory (Tanaka, 1982).

2.3.1 Fuzzy modeling in medical

Hata et al. (1998) described usefulness of fuzzy logic techniques for medical images segmentation. The fuzzy logic techniques included fuzzy information granulation, fuzzy inference and fuzzy cluster identification. The function of fuzzy information granulation is to introduce the powerful scheme to find the whole brain region in medical data. Furthermore, fuzzy inference succeeded to segment the brain region into left and right hemisphere, cerebellum and brain stem and lastly, the fuzzy
clusters identification is to determine the clusters into blood vessel or other tissues in magnetic resonance angiography (MRA) image.

Liu et al. (2005) studied on application of fuzzy c-means and fuzzy integral theories to a proposed two-step machinery fault diagnosis model. They combined fuzzy measures and fuzzy integral and applied to diagnose results into a consensus final decision. Vibration signals from rolling element bearings were used to validate the method. The results showed that using fuzzy c-means and fuzzy integral techniques improved the diagnosis accuracy and reduced the computation load.

Baig et al. (2011) identified a model to control strategies in the medical field with the aim to diagnosis the diseases. They identified fuzzy logic medical diagnosis control system is suitable to control diagnosis of hemorrhage and brain tumor. This model used design and simulation as control system. As general results, fuzzy logic medical diagnosis control system developed a control system to enhance the efficiency to diagnose a disease related to human brain.

Ibrahim et al. (2011) founded Optimized Fuzzy Control For Natural Trajectory Based Fes-Swinging Motion. The study was to get the swinging of legs during movement happened. Fuzzy trajectory tracking control of swinging motion with genetic algorithm was applied. After the analysis was done, the results showed its effectiveness in controlling swinging motion in the simulations as well as in the practical environment.

Rajeswari and Vaithiyanathan (2011) discussed the variation of preliminary inquiry information obtained from patients of a diabetic and research center using fuzzy relation based model. The parameters were modeled using a fuzzy approach. After normalization, it was classified by using artificial neural networks as ‘Close to Type 2 diabetic’ or not. This result in this study, collected in stage I are not enough to classify type 2 diabetes of patients and further stages are required for improvement.

Rusiman Saifullah et al. (2012) proposed two models of regression which are data modification by using analytical hierarchy process (AHP) technique and fuzzy c-mean (FCM) model. The objective of the study was to study the scale of health at an intensive care unit (ICU) ward using the AHP, FCM model and FCRM models. These models have been compared using MSE values. After comparing the four
cases, it was found that the AHP technique appeared to be the best model with the lowest mean square error (MSE). Moreover, AHP technique can increase the accuracy of modeling prediction.

Elizabeth and Sujatha (2013) proposed a fuzzy medical diagnosis network for decision making in medical diagnosis. They applied medical diagnosis under fuzzy environment using triangular fuzzy number and membership of matrix. Function of the model is to determine the contradiction of transitivity of the ranking. This form of transitive ranking can be accommodated by means of relativity function. This study able to help patients with accurate medical diagnostic process and decision making process.

Yilmaz and Ayan (2013) studied on cancer risk analysis by adopting fuzzy logic approach and performance status of the model. This study applied the fuzzy logic created by Zadeh in 1965 on cancer risk. Three cancer risks were selected which are breast cancer, lung cancer and colon cancer. Analyses were carried out with proposed method of fuzzy logic, rating formula and fuzzification, rule processing unit and defuzzification. This study helps people to measure the developing cancer and simultaneously, the rate of suffering from cancer can be decreased.

El-Melegy and Mokhtar (2014) proposed a new fuzzy approach for the automatic segmentation of normal and pathological brain magnetic resonance imaging (MRI). The model is fuzzy c-means (FCM) algorithm. Constrained minimization function using the lag range multiplier was constructed. There are 2 steps to approach FCM which are solution estimation and prior information guided solution. FCM algorithm boosts the performance under noise and unexpected data acquisition conditions. The results of FCM were considerable better segmentation with accuracy, robustness against noise, and faster response compared to several well-known fuzzy and non-fuzzy techniques.

Udapa et al. (2014) discussed a body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. This study applied Iterative Relative Fuzzy Connectedness (IRFC) delineation algorithm to build a general automatic anatomy recognition (AAR) system that is not tied to any specific organ system, body region or image modality. This model containing five steps which gathered image data for both building models and testing the AAR algorithms from patient image sets existing in human health system. Followed by formulating
precise definitions of each body region and organ and delineating them following these definitions, building hierarchical fuzzy anatomy models of organs for each body region, recognizing and locating organs in given images by employing the hierarchical models and delineating the organs following the hierarchy. The results showed that fuzzy model-based IRFC are effective concepts which allowed to demonstrate the feasibility of a general AAR system that works in different body regions on a various organs and on different modalities.

2.3.2 Fuzzy linear regression models

In this study, there are four models of fuzzy linear regression applied to analyze colorectal cancer stages in Malaysia.

Tanaka et al. (1982) proposed linear regression analysis with fuzzy model. This structure is represented fuzzy linear function by fuzzy sets. Fuzzy linear functions proposed by Zadeh’s extension principle. They used input and output data concerning house price as the data and using fuzzy linear regression model. They concluded that fuzzy linear function to the linear regression model was an established method for vague phenomena.

Ni (2005) proposed a new linear programming for fuzzy linear regression extension by Tanaka (1982). The purpose of this research is to revisit the ongoing research work that people have already done on this issue and to develop new models related to fuzzy data correlation and regression. The research used Intel Corp. 2003 stock price as data and the author concluded that fuzzy regression models could be applied in short term stock price prediction. Moreover, he found that fuzzy regression gives more consistent results, since it successfully modeled the inherent vagueness which exists in the application by formulated form.

Chung (2012) proposed fuzzy linear regression benchmarking model under fuzzy environment. Benchmarking systems had been developed to conduct benchmarking process for the energy efficiency of buildings. However, not all benchmarking systems can develop it because of the different methods in developing systems. An approach for benchmarking the energy efficiency of buildings, statistical regression analysis had been used. At present, the method was used in a crisp input and output problem. By adopting the standardization of the input data, the result
showed that fuzzy linear regression model can be used to build a benchmarking model.

Zolfaghari et al. (2014) proposed a new linear programming fuzzy linear regression with symmetric parameter. This linear programming proposed to determine relationship between overall acceptance and sensory characteristics (aroma, surface color, porosity, hardness, oiliness, and flavor) of 36 different types of fried donut. Modeling was done by assuming that independent variables are crisp and coefficients are triangular fuzzy number. This study proposed two models which are fuzzy linear regression with symmetric parameter and fuzzy linear regression with asymmetric parameter. The results showed that symmetrical fuzzy parameter provided the best fitting of sensory data.

2.4 Variables review

Zhang (2004) studied the patients’ age for a modest positive correlation with incident colorectal cancer. The older the people age may just a generational effect of colorectal cancer. While some study have found that race also has significant correlation with incident colorectal cancer. This study used 14 factors as predictors by logistic regression which are age, gender, race, stage of colorectal cancer, tumour size and many more.

Wendy & Radzi (2008) proposed national cancer patient registry of colorectal cancer. In Malaysia, no such comprehensive report on all aspects of colorectal cancer exists. This issue is raised in Ministry of Health in registry of colorectal cancer field and 197 cases of colorectal cancers were reported in the database between October 2007 and February 2008. The collected data were demographic details, family history, pathology staging, follow-up details, therapy offered, presenting symptoms and concomitant medical problems. This study analyzed using multi centre project.

Xu et al. (2009) applied comparison of adaBoost and logistic regression. AdaBoost call a base learning algorithm repeatedly based on the error. The aim of this study is to detect colorectal cancer patients with synchronous liver metastasis. There were several variables used in this study which is part of analysis. The data were taken from Department of Colorectal Cancer at Shanghai Tumor Hospital from January 2000 to June 2005. The included variables were thinner stool, hard stool,
diarrhea, abdominal mass, rectal, ascending colon, sigmoid colon, rectum and many more.

Varela et al. (2010) studied the factors and causes of colorectal cancer incident. The study was conducted to understand the factors related to colorectal cancer (CRC) screening among urban Hispanics. The included factors were country of origin and race, family history of CRC and relationship with colorectal cancer. This study used descriptive analysis to achieve the goal of the study.

Stefani et al. (2011) in the period of 1996-2004 case control colorectal cancer was conducted in Montevideo, Uruguay. Recently, the case controls have frequency matched for age, sex, family history and residence. All cases were conducted from the same general hospitals. This study was conducted to determine the risk of factors on colorectal cancer analysis in Uruguay using Kaiser Method.

Roslani et al. (2012) found that blood tool is one of the main factor causes colorectal cancer in human body. The screening for colorectal neoplasias with fecal occult blood test: false-positive impact of non-dietary restriction. The aimed of this study to evaluate and compare the validity of guaiac-based fecal occult blood test (gFOBT) and fecal immunochemical test (FIT) in Asian population. gFOBT has many advantages compared to FIT, in terms of the cost of the treatment. There were two categories to test blood tool using gFOBT which are structural exam and fecal test. The findings showed that the gFOBT test was positive.

Phillips et al. (2013) studied on examining the differences in rates of initiating colorectal cancer screening across age. This study found that the screening colorectal cancer remain low among adults age 50 and older in the United States. However, this study used logistic regression as solutions because that model widely applied in many biomedical fields. The rest of various factors are sex, education, race, body mass index, colorectal screening and others.

2.5 Summary

Table 2.1 is a summary of the literature review on colorectal cancer in the previous studies. It shows that descriptive analysis, logistic regression and linear regression are the most common methods used to analyse the colorectal cancer. Other than that, fuzzy linear regression is not widely used in colorectal cancer data analysis.
However, fuzzy linear regression model is also often used in medical research at the present time.

In the past research, many methods used colorectal cancer analysis in general hospitals which are descriptive analysis, logistic regression and linear regression. Only some researchers used fuzzy model in solving the colorectal cancer cases but not fuzzy linear regression. Therefore, fuzzy linear regression will be one of the best model and tends to contribute in the study to detect tumor size of colorectal cancer by factors and symptoms.

There are several different between this study and previous study such as this study applied all stages of colorectal cancer (I, II, III and IV) and using all the fuzzy linear regression models as methodology. The use of tumor size in colorectal cancer has a great potential of methods to enhance predicting tumor size by factors and symptoms of colorectal cancer; therefore the fuzzy linear regressions should be develop based on mean square error and root mean square error in colorectal cancer.

Table 2.2 shows the summary of fuzzy model in medical field. Fuzzy model is a general model of fuzzy which was used by any researcher in medical analysis. Furthermore, linear regression included multiple linear regression, logistic regression and fuzzy regression.
REFERENCES


Frieden. (2002). Physical Activity, Dietary Fat and Colorectal Cancer. PhD Thesis Martina Perše University of Ljubljana, Faculty of Medicine, Institute of Pathology, MEC, Slovenia.


National Cancer Registry, Ministry of Health Malaysia. (2006). *Malaysian Cancer Statistics – Data and Figure Peninsular Malaysia.*

National Cancer Registry, Ministry of Health Malaysia. (2010). *Malaysian Cancer Statistics – Data and Figure Peninsular Malaysia.*


