UNIVERSITI MALAYSIA SARAWAK

UNIMAS

Faculty of Computer Science and Information Technology

SURVIVAL ANALYSIS IN MEDICAL DATASETS

Joyce Goh Chui Wen

Bachelor of Computer Science with Honors

(Computational Science)

2020

SURVIVAL ANALYSIS IN MEDICAL DATASETS

JOYCE GOH CHUI WEN

This project is submitted in partial fulfillment of the requirements for the degree of

Bachelor of Computer Science with Honors

(Computational Science)

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2020

ANALISIS SURVIVAL DALAM DATASET PERUBATAN

JOYCE GOH CHUI WEN

Projek ini merupakan salah satu keperluan untuk

Ijazah Sarjana Muda Sains Komputer dan Teknologi Maklumat

(Sains Komputan)

Fakulti Sains Komputer dan Teknologi Maklumat

UNIVERSITI MALAYSIA SARAWAK

2020

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE SURVIVAL ANALYSIS IN MEDICAL DATASETS

ACADEMIC SESSION: 2019/2020

JOYCE GOH CHUI WEN
(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

□ CONFIDENTIAL (Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)

□ RESTRICTED (Contains restricted information as dictated by the body or organization where the research was conducted)

☑ UNRESTRICTED

Validated by

Phang

(AUTHOR'S SIGNATURE) (SUPERVISOR'S SIGNATURE)

Permanent Address

1J, JALAN PERMAI TIMUR 6
96000, SIBU
SARAWAK

Date: 03.08.2020 Date: 03.08.2020

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

Declaration

I hereby declare that this research together with all of its content is none other than that of my own work, with consideration of the exception of research based information and relative materials that were adapted and extracted from other resources, which have evidently been quoted or stated respectively.

Signed,

……………………………………………

JOYCE GOH CHUI WEN

Faculty of Computer Science and Information Technology 07/12/2019

Universiti Malaysia Sarawak

Acknowledgement

First and foremost, I would like to express my great appreciation to Faculty of Computer Science and Information Technology (UNIMAS) for giving me such precious chance to undertake and complete my Final Year Project (FYP). It was a great chance for me to learn and enhance my IT knowledge.

Besides, a very sincere and special thanks to my respected supervisor, Dr Phang Piau, who always shares his expertise, advice and valuable guidance for me to complete my FYP. Without his help, I might not be able to complete my FYP within the given period.

Last but not least, I would like to express my thanks to my families and friends for their encouragement and support all the time through my FYP period. Their support was motivating and inspiring me to complete this project.

Abstract

This project is studied about the survival analysis in several medical datasets. Survival analysis is a method for data analysis in which the outcomes indicate the time to the occurrence of an event of interest. By time, it can be years, month, weeks or days from the beginning of follow-up of an individual until an event occurs; alternatively, it can refer to the age of an individual when an event occurs. In medical studies, time to death is the event of interest. This study is based on 11627 observations comprising of 6605 females and 5022 males. The age of the observations is in the range of 32-81 years old. The dataset is retrieved from Framingham Heart Study. The data was collected during three examination periods, approximately 6 years apart. The aim of this research is to explore the selected medical datasets by using data visualization techniques for better insight and manipulation tools in R Studio. The Kaplan-Meier plot was used to study the general pattern of survival which showed the survival rate of the patients. Cox regression was used to study the regression coefficient, hazard ratio, standard error, statistical significance, p, Likelihood ratio test, and p-value. The result shows that gender, age, and blood pressure are found impacting the disease development.

Abstrak

Projek ini dikaji mengenai analisis survival dalam beberapa kumpulan data perubatan. Analisis survival adalah satu kaedah untuk analisis data di mana hasil menunjukkan masa untuk berlakunya kejadian yang menarik. Menjelang masa, ia boleh bertahun-tahun, bulan, minggu atau hari dari permulaan tindak lanjut seseorang sehingga peristiwa berlaku; Sebagai alternatif, ia boleh merujuk kepada umur individu apabila sesuatu peristiwa berlaku. Dalam kajian perubatan, masa kematian adalah peristiwa yang menarik. Kajian ini berdasarkan kepada 11627 pemerhatian yang terdiri daripada 6605 wanita dan 5022 lelaki. Umur pemerhatian adalah dalam lingkungan umur 32-81 tahun. Dataset diambil dari Kajian Hati Framingham. Data dikumpul selama tiga tempoh pemeriksaan, lebih kurang 6 tahun. Tujuan penyelidikan ini adalah untuk meneroka dataset perubatan terpilih dengan menggunakan teknik visualisasi data untuk alat wawasan dan manipulasi yang lebih baik di R Studio. Plot Kaplan-Meier digunakan untuk mengkaji corak umum kelangsungan hidup yang menunjukkan kadar survival pesakit. Regresi kox digunakan untuk mengkaji koefisien regresi, nisbah bahaya, ralat piawai, kepentingan statistik, p, ujian nisbah kebolehan, dan nilai p. Hasilnya menunjukkan bahawa jantina, usia, dan tekanan darah didapati mempengaruhi perkembangan penyakit.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

Chapter 1 Introduction

This chapter will cover the introduction/background of study, problem statement, scope, aims and objectives, brief methodology, significant of project, project schedule and expected outcome.

1.1 Introduction/Background

Heart disease is a class of diseases that involves heart and blood vessels. It is also a common type of non-communicable disease in which the disease is not transmissible directly from one person to another. It includes stroke, hypertensive heart disease, heart failure, and etc.

According to the World Health Organization (WHO), an estimated 17.7 million people worldwide died from heart disease in 2015, representing 31 percent of all global death (Bernama, 2018). Unhealthy lifestyle, as well as poor diet, obesity, smoking, and high sugar intake may lead to develop heart disease. However, the chances of having heart disease can be reduced through the lifestyle changes. While there are no guarantees that healthy lifestyles can totally prevent the developing of heart disease, the changes will definitely improve the health in other ways, such as improving physical and mental health. In fact, there are still some uncontrolled risk factors such as age, and family history of heart disease. Therefore, it is important to understand the prevalence of risk factors for heart disease.

Survival analysis is a method for data analysis in which the outcomes indicate the time to the occurrence of an event of interest. By time, it can be years, month, weeks or days from the beginning of follow-up of an individual until an event occurs; alternatively, it can refer to the age of an individual when an event occurs. By event, it can include death, existence of disease, and etc that may happen to an individual. The survival time refers to the time in years until a person suffers the heart disease (Despa, n.d.).

One of the most popular regression techniques for survival analysis is Kaplan-Meier estimator, also known as the product-limit estimator. In medical research, it is used to measure the fraction of patients living for a certain amount of time after treatment. The time is starting from a defined point to the occurrence of a given event (i.e., death) is called as survival time, while the analysis of group data is called as the survival analysis (Kishore, Goel, & Khanna, 2010).

Cox regression which can also be described as proportional hazards regression is a method to investigate the effect of several variables upon the time a particular event takes to happen ("Cox Regression (Proportional Hazards, Hazard Ratio) - StatsDirect," n.d.). The measured effect of this model is hazard rate, which is the probability of suffering the event of interest.

Therefore, the regression techniques for survival analysis will be used to formulate the probability risks to suffer from heart disease in this study. In addition, data visualization techniques and manipulation tools using R programming will be used to have a better understanding of the selected medical dataset.

1.2 Problem Statement/Research Problem

Heart disease is the leading cause of death for both men and women. However, it is difficult to identify heart disease due to several contributory risk factors. This project aims to study the survival analysis in heart attack by using R programming. Survival analysis is a method of making the prediction at various points in time, which means, while other predictive models can predict whether an event will occur, the survival analysis predicts whether the event will happen at the specified time. By using data visualization technique and manipulation tools in R Studio, the probability risks of developing heart disease of patients that have experienced a heart disease and those who do not in accordance with the covariates: age, gender, diabetes, unhealthy lifestyle, or family history can be found out. With the probability of risks to suffer from heart disease, the rate of survival might be increasing as people can manage well with the early detection, prevention, and treatment. Hence, people may increase the awareness to suffer from heart disease, unless the uncontrolled factors from age and family history.

1.3 Scope

This study is based on 11627 observations comprising of 6605 females and 5022 males. The age of the observations is in the range of 32-81 years old. The dataset is retrieved from Framingham Heart Study. The data was collected during three examination periods, approximately 6 years apart.

1.4 Aims and Objectives

(1) To explore selected medical datasets using data visualization and manipulation tools using R programming.

(2) To carry out the survival analysis for predicting the risk of heart disease.

1.5 Brief Methodology

This project will employ statistical modeling process methodology. This study aims to predict the survival time of the observers whose data are collected on the medical datasets. In addition, the risk to suffer from heart disease is predicted because it is difficult to identify heart disease due to several contributory risk factors. The dataset was retrieved from Framingham Heart Study. The dataset is then saved in CSV (comma delimited) format file.

The methodology chosen should be related to survival analysis. Therefore, a data visualization technique is used as a guideline for this proposal to explore the selected medical datasets. Besides, several manipulation tools in R Studio will be used to carry out the survival analysis to predict the risk of heart disease.

1.6 Significance of Project

The probability of getting the risk of heart disease can be formulated by using regression techniques for survival analysis. Besides, data visualization techniques and manipulation tools in R Studio are used to have a better understanding of the selected medical datasets.

1.7 Project Schedule

The project schedule includes Final Year Project and shows in Appendix A, Figure A.1.

1.8 Expected Outcome

The expected outcome of this project is exploration by using data visualization for gaining a better understanding on the selected heart disease dataset. In addition, a statistical analysis with incorporate survival analysis for selected medical dataset will be generated by using the manipulation tools using R programming.

Chapter 2 Literature Review

This chapter will cover the brief introduction to survival analysis and some of its terminologies which includes time to event, censoring, survival (and hazard) function, and proportional hazards model. Besides, this chapter also includes comparison of some statistical tools which are Minitab, SPSS, SAS, and decided statistical tools use in this research, R studio. Besides, other statistical research on survival risk prediction also been studied for this chapter. The review is the majority based on the functionalities, advantages, and disadvantages of these tools. The comparison of features, strength, and weakness among the existing and proposed tools was studied and analyzed.

2.1 Survival analysis

Survival analysis is used to analyze the data in which involving times to some event of interest. The events can be death, the occurrence of a disease, marriage, divorce, and other. On the other hand, survival analysis is also a method to estimate the lifespan of a particular population under study.

In survival analysis, the subjects are followed over a specified time and the focus is on the time at which the event of interest happens. The unique features of time to event variables include the times to event are always positive and the distributions are often skewed.

According to Kartsonaki (2016), the starting point of time must be specified so that everyone is as equal as possible. To give an instance, if the survival time of patients with a particular type of cancer is being studied, the origin of time could be chosen as the starting point of diagnosis of that particular type of cancer. It is also important that the endpoint or event of interest should be specified appropriately so that the determined time is clearly defined.

Some of the observations who may withdraw from the study, or they may have some events such as in the above example, dying due to an accident, which is not a part of the target endpoint may not be observed. Besides, while the study was completed at some point in time, yet the individual has not yet had an event, their event time will not be observed. These incomplete observations need to be handled in different ways, which called censoring.

Censoring is an important issue in survival analysis which represents a particular type of missing data (Despa, n.d.). The most commonly encountered type of censoring is right censoring. It occurs when the observation has dropped out from the study before an event happens. Often, the study ends at a specific time, and some individuals have not yet had an event at the end of the study. Another type of censoring in survival analysis is left censoring. It is a situation in which the individual is known to have the event occurred before a certain time, but may be at any time before the specific time. It can occur when a person's true survival time is less than or equal to that person's observed survival time. The interval censoring happens when the follow-up period is not exact or continuous. It occurs when failure is only known to have occurred during an interval.

The survival function can be plotted by using a set of data. Kaplan-Meier curve is a non-parametric estimator for survival function to do prediction. In medical research, it is used to measure the fraction of patients living for a certain amount of time after treatment. The time is starting from a defined point to the occurrence of a given event, for example, death is called as survival time and the analysis of group of data as survival analysis (Kishore et al., 2010). The Kaplan-Meier curve is also used to compare the survival times of two or more group.

Moreover, another popular regression model in survival analysis is Cox proportional hazards regression. Cox regression is a semi-parametric model which is commonly used to investigate the effect of several variables against the specified time on particular events to happen ("Cox

Regression (Proportional Hazards, Hazard Ratio) - StatsDirect," n.d.). It provides the ease to interpret the information that regards the relationship of the hazard function to predictors. For example, in the medical research, Cox proportional hazards model is used to find out which variable has the most important impact on the survival time of a patient ("Cox proportional hazards models | Statistical Software for Excel," n.d.).

2.2 Other Statistical Research on Survival Risk Prediction

In many clinical setting, statistical models are being developed to predict the risk of disease or other adverse events. These models are designed to help patients and doctors for making a wise decision. The value of models would be the guidance of the decisions. Common types of models include the logistic regression model, Cox proportional hazards, and classification trees.

2.2.1 Breast Cancer by Tice JA

Tice et al. (2008) aims to develop and validate the breast cancer risk prediction model includes breast density has studied a model to predict the breast cancer risk that only includes age as the predictor. For a woman in a certain age, the resulting risk estimate is simply the percentage of women in her age group who will develop breast cancer. The predicted risk will be changed if more information is added in the model. For example, if family history information is included, a woman's predicted risk will be the proportion of women age with her family history that will develop breast cancer.

The data is obtained from 1,095,484 women who had no previous diagnosis of breast cancer. They were asked to undergo mammography, and morbidity from Surveillance, Epidemiology, and End Results (SEER) to do the model evaluation for predicting the breast cancer risk. Cox proportional hazard model is used and the attributes included age, race/ethnicity, family history, and history of breast biopsy. This study is focused on the effect of adding breast density information to the model.

According to Tice et al., (2008), the results of this study show that 14,766 women have diagnosed the invasive breast cancer during 5.3 years of follow-up. The limitation of the model has only ability to discriminate between women who will develop breast cancer and who will not. However, the accuracy still needs to be further to evaluate in independent populations.

2.2.2 Ovarian Cancer by K Li

Ovarian cancer has a high mortality rate due to late diagnosis. However, epidemiological risk prediction models might help to identify women at increased risk who may benefit from targeted preventive measures. The study is aimed to build an ovarian cancer risk prediction model for women in Western Europe.

Li et al. (2015) built an ovarian cancer risk prediction model with epidemiologic risk factors from 202,206 women in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. The data is based on epidemiological questionnaire data from European women aged 45 years and over. The risk factors are included: status, age, duration of HRT, duration of OC use, unilateral ovariectomy, number of FTPs and BMI were selected as major predictors.

The result of this study shows that there are 791 primary ovarian cancers were diagnosed after a median follow up time of 11.7 years, and 324 of them were diagnosed within the first 5 years follow up. Median age at recruitment was 52.4 years which the range is between 45 – 78 years old and baseline BMI was on average 24.5 kg m^{-2}. Hence, it can suggest that, older age and higher BMI were associated with an increased ovarian cancer risk (Li et al., 2015). The informative biomarkers should consider in the future studies may improve the predictive ability of the model.

2.3 Other Statistical Software used for Data Analysis

2.3.1 Minitab

Minitab is statistical software which automates calculations and the creations of graphs allow users to have more focusing on the data analysis and interpretation of results ("Minitab Statistical, Data analysis & Process Improvement software," n.d.). It is used by different companies to graph and to analyze their business data. It is a user-friendly tool which not only used as a business tool but also used by colleges and universities to teach statistics and also data analysis. However, it has poor compatibility when compared with other statistics programs, which is, difficulty in making file to import. It is not a good choice when comes for organizations that may need to do combination data from multiple sources.

2.3.2 SPSS

Statistical Package for the Social Science (SPSS) which currently named IBM SPSS Statistics is a software package that used for the analysis of statistical data (Rouse, 2018). It is commonly used in healthcare, marketing, and education research. In addition to statistical analysis, the features of the base software also include data management and data documentation. SPSS is easy to use for importing data or extracting data in a different file format such as comma-separated values (CSV). Besides, it can be used to process multiple data types and varieties. It has a graphical output that is less ideal in terms of how the data is presented, so it might be difficult to use the output. It is important to make sure that a careful analysis of results because sometimes, the statistical results can be misinterpreted. SPSS is a commercial product so it would be cost a reasonable amount to purchase ("SPSS and Data Analysis 'SPSS Statistics Software' | More Clarified," n.d.).

2.3.3 SAS

Statistical Analysis System (SAS) is an important analytics tool extensively used by data scientists and data analytics. SAS is preferred over R programming language and Python. SAS is a programming language that used to read the data from spreadsheets and databases. It is usually used for financial analytics. SAS consists of various tools such as graphs, plots, and highly versatile libraries. SAS also provides a graphical point-and-click user interface for non-technical users. However, it cannot do the file sharing with another user who does not use SAS. It also lacks of graphic representation and difficult on text mining (Team, 2019).

2.4 R studio

R studio is an open-source environment which is known for its simplicity and efficiency (Malik, 2019). It can be categorized as powerful data analysis software that leverages the power of R to facilitate its use in the field of data science and statistics. It provides a complete environment for data scientists to work in data-driven industries such as healthcare, finance, technology, and other scientific research. This means that all the data types, statistical model, data processing, and charts required are based on the strong foundation of R. R programming language is an important tool for development in the fields of numerical analysis and machine learning (Krill, 2015). In the basic Graphic User Interface (GUI) in R studio, the stored objects and data will always be listed in the "Global Environment" window and just click on the data sets directly to open and view as spreadsheets. It makes convenient for users instead of memorizing where the data sets stored. R packages that contain reproducible R code are the fundamental units developed by R StudioInc. and its users. In this final year project, packages that would be used include survival, tidyverse, and ggplot2. Tidyverse was created by the great Hadley Wickham and his team with the objective of providing all the utilities to clean and work with data (Analystics Vidhya, 2019). The packages under Tidyverse help to perform and interact with data. Whereas, ggplot2 is a data visualization package that allows user to add, remove or alter components in a plot.

2.5 Data Visualization

Data Visualization is an art of turning data into insights that can be easily interpreted. It involves representing data in graphical or pictorial form which makes the information easy to understand (Adam Heitzman, 2019). It helps in explaining the facts and determining the courses of action. In the world of big data, it is difficult for users to analyze massive amounts of data, information, and make data-driven decisions. Thus, data visualization technique plays an important part in data analytics and helps to interpret big data in a real-time structure by leveraging complex digital or factual data sets. Data visualization is also regarded as information visualization or scientific visualization. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to have better insight and understanding of trends, outliers, and patterns in data. Therefore, decision-makers are able to comprehend the information, discover the patterns, and make the decision easily.