CHAPTER ONE

INTRODUCTION

Survival analysis is concerned with time-to-event data. The survival time is measured from a well-defined time origin until the occurrence of some particular event. The time origin corresponds to the recruitment time of the individual into the experimental study. The occurrence of event may coincide with the diagnosis of a particular condition, or the commencement of a treatment regime. Event is usually referred to as "death" or "failure". Thus, time from recruitment to the death is referred as the survival time of the patient. Event can also be non-fatal, such as relief of pain, or recurrence of symptom and disease.

In survival data, two types of observations are considered; censored and uncensored. Uncensored observation refers to individual whose event is observed in the study. In contrast, there are individuals with unknown status of event. For example, patient is still alive when the study ends, or patient is lost to follow-up, or withdraws for reasons unrelated to the study. Thus, the information available is only the survival experience of the patient and the last date of patient known to be alive. Such observations are referred as censored observations.

The simplest method of describing survival data is based on the non-parametric approach. The main aim is to estimate the survival probability of an individual at a given time *t*. Several non-parametric methods are available for this purpose. The common methods used are the life-table and the Kaplan Meier methods (Kaplan and Meier (1958)). The estimate represents the survival probability of an individual from a well defined time to sometime beyond the time origin. On the other hand, the Nelson-

Aalen method is usually used to estimate the hazard function (see Aalen (1978), Altshuler (1970) and Nelson (1972)). The hazard estimate represents the probability that an individual dies at a particular time.

In the medical field, the supplementary information contained in explanatory variables of individuals is recorded. Explanatory variables can be categorized into three groups; demographic variable such as the age, race and gender of the patients, physiological variable such as serum haemoglobin level and heart rate and, lastly, factor associated with the lifestyle such as smoking history and dietary habit. The explanatory variables may have impact on the survival time of the patient.

In order to explore the relationship between the survival experience of individuals and explanatory variables, the data can be modeled using various types of survival model. Through modeling, a subset of important explanatory variables can be determined. Thus, the impact of these variables on the survival probability or the hazard of individuals can be studied. The most widely used model in survival analysis is the proportional hazard model (PHM). This model assumes that the hazard between groups is proportional for all time *t*. This assumption is known as the proportional hazard assumption (PHA). Various graphical and numerical tools are available to investigate the assumption (see Andersen (1982), Breslow *et al*. (1984), Arjas (1988) and Quantin *et al.* (1996)).

A probability distribution plays an important role in survival data analysis. For example, if the hazard is shown to follow Weibull distribution, then the appropriate parametric PHM is the Weibull PHM. However in many cases, we may not be able to identify the distribution of the survival data. In such situations, we usually use the Cox

PHM proposed by Cox (1972) (see Marubini and Valsecchi (1995), Kleinbaum (1996), Hosmer and Lemeshow (1999) and Collet (2003)). The Cox PHM also requires the PHA to be satisfied.

An essential part of any statistical modeling is to investigate the adequacy of the model. In our case, if the assumption of the model is not met, then the fitted model is not suitable for the data. The diagnostic checking may be carried out in several stages. The most important stage is to verify the validity of PHA. Then, the functional form of the experimental variables should be investigated given that the effect of other covariates is already accounted for. It is important as well to screen for the presence of outliers and influential observations in the data. In this study, our interest is on improving the procedure of detecting outliers and developing procedures of identifying influential observations in the Cox PHM.

1.1 Survival Analysis

Survival data has two main characteristics; the presence of censored observation and the non-symmetrical distribution of survival time. Censored observations refer to observations whose survival experience is partially known. One of the reasons for this is that the person studied is still alive when the data is evaluated, and thus his complete lifetime is not known at the end of the study. There are three types of censoring: right censoring, left censoring, and interval censoring. The right censored survival time is less than the actual, but unknown survival time. For example, the survival time of leukemia patients is measured until the remission time. On the other hand, the left censored survival time is encountered when the event occurs before being detected at some time later. For example, we will never know the actual date of a patient being infected by HIV, but we know the detection time during the hospital visit. In general, the left censored survival time is longer than the actual one. Left censoring occurs far less commonly than right censoring. Meanwhile, the interval censoring occurs when individuals are known to have experienced an event within an interval of time.

In summarising survival data, two functions of central interest are the survival function and the hazard function. Let the actual survival time *t* of an individual be regarded as the value of a variable *T*, which can take non-negative value. Survival function $S(t)$ is defined to be the probability that the survival time is greater than or equal to *t* given by

$$S(t) = P(T \geq t) = 1 - F(t) \qquad (1.1)$$

where *F*(*t*) is the cumulative distribution function.

Meanwhile, the hazard function of *T* is defined to be the probability that an individual dies at time *t*, conditional on he or she is surviving to that time. So, the hazard function is then the limiting value of $P(t \leq T < t + \delta t / T \geq t)$ divided by δt, as δt tends to zero, given by

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \right\}. \qquad (1.2).$$

Equation (1.2) is also called the instantaneous death rate and can be simplified as

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t) \cap P(T \geq t)}{\delta t \;\, P(T \geq t)} \right\} \\ &= \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{P(T \geq t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

Consequently, it can be shown that

$$h(t) = -\frac{d}{dt}\left(\ln S(t)\right)$$
$$S(t) = \exp\{-H(t)\} \qquad (1.3)$$

where $H(t) = \int_0^t h(u)$ is the integrated (cumulative) hazard, which also can be written as

$$H(t) = -\log S(t) \qquad (1.4).$$

From the equations (1.1) to (1.4), it can be seen that the hazard and survival functions are related to each other. For example, consider a constant hazard function for some specific values λ given by $h(t) = \lambda$. It is easily shown that the respective survival function is given by $S(t) = \exp(-\lambda t)$ with the density $f(t) = \lambda \exp(-\lambda t)$. In this case, the survival times are said to follow exponential distribution. In other cases, we might need to consider other types of distribution if certain condition is satisfied. For instance, if the hazard function increases or decreases monotonically with increasing survival time, we should consider the Weibull distribution to describe the survival data. In fact, this is the most widely used distribution in the medical fraternity.

1.2 Survival Model

The proportional hazard model (PHM) is the most common model used to model the survival data. This model assumes that the hazard of death between different groups is proportional at any time *t*. The general model is considered as the hazard of death at a particular time *t* given the values $x_1, x_2,..., x_p$ of *p* explanatory variables $X_1, X_2 \cdots X_p$. Let $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots x_{ip})'$ and $h_0(t)$ be the hazard function for an individual with $\boldsymbol{x}_i = \mathbf{0}$. $h_0(t)$ is called the baseline hazard function. Then the hazard for the *i*th individual can be expressed as

$$h_i(t) = \psi(\boldsymbol{x}_i)h_0(t) \tag{1.9}$$

where $\psi(\boldsymbol{x}_i)$ is a function of the values of the vector of explanatory variables of the *i*th individual. The function $\psi(.)$ can be interpreted as the hazard at time t for an individual whose vector of explanatory variables is $\boldsymbol{x}_i$, relative to the hazard for an individual for whom $\boldsymbol{x}_i = \boldsymbol{0}$. $\psi(\boldsymbol{x}_i)$ is known as hazard ratio.

Since $\psi(\boldsymbol{x}_i) > 0$, the hazard for the *i*th individual can be written as

$$h_i(t) = \exp(\eta_i)h_0(t) \tag{1.10}$$

where

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \tag{1.11}$$

is called the linear component of the model, or the risk score, or the prognostic index for the *i*th individual. Model (1.10) can be further rewritten as a linear model for the logarithm of the hazard ratio

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\,. \qquad (1.12).$$

This model is better known as Cox PHM (see Cox (1972) and Collet (1994)). The model is considered as a semi-parametric model since it does not assume any underlying distribution for the survival times.

1.3 Outlier and Influential Observation

The problem of outliers in univariate and regression data has been discussed in the literature for more than a century. Barnett and Lewis (1984) defined outlier as "an observation which may appear to be inconsistent with other observations in the data set". Earlier, Beckman and Cook (1983) defined outlier as the observation that is not a realization of a target distribution. According to Hawkins (1980), "an outlier is an observation that deviates so much from other observations so as to arouse suspicion that it was generated by different mechanism". Meanwhile, in survival analysis the outlier is defined slightly differently from linear regression problems. This is because the dependent variables in survival models contain the information on the survival times and status of patients in the study.

It is well known that the existence of outliers may alter the parameter estimation of fitted models in all areas of research. Belsley *et al.* (1980), Barnett and Lewis (1984) and Montgomery and Peck (1992) had discussed the occurrence of outliers in linear regression problems. In survival analysis, the outlier can also affect the parameter estimation of the models, alter the respective hazard ratios, and may change the selected model.

In survival data, many authors have tried to give specific meaning to the outlier due to the special features of the data. Collet (2003) referred outlier in survival as an individual who has extremely long survival time, but the values of the explanatory variables suggested the individual should have died earlier, and vice versa. Therneau *et al.* (1990) and Nardi and Schemper (1999) associated outlier to individuals who "died too soon" or "lived too long", while Maller and Zhou (1994) identified outlier as individual who is already "immune" or "cured" after being released from prison. If this individual affects the inferences made based on the model, then this individual is identified as an influential observation.

Usually, outliers are screened using the deviance residuals proposed by Therneau *et al.* (1990). However, Flemming and Harrington (1991) pointed out that the deviance residuals have no reference sampling distribution and the approximation by a standard normal distribution appears non-satisfactory even without censored observations in the data set. Later, Nardi and Schemper (1999) proposed new types of residuals to overcome this problem. It is claimed that procedures based on these new residuals perform better in correctly identifying outliers. The residuals are the log-odds and normal deviate residuals.

Beside the problem of outlier detection, there is also a need to develop a procedure to identify influential observation. In regression problem, Belsley *et al.* (1980), Hadi (1992) and Imon (2005) had successfully identified influence observations in linear regression problems using the leave-one-method. On the other hand, similar method has also been proposed to identify influential observation in survival problems. Beside the standard exact delta-beta procedure, Cain and Lange (1984) and Reid and Crépeau (1985) used an influence function (IF) in investigating the problem, while Storer and Crowley (1985) proposed the augmented approach (AUG). Wang *et al.* (2006) compared all the approaches above and showed that the AUG approach clearly outperforms the others.

Meanwhile, Atkinson and Riani (2000) had proposed an alternative method called forward search method to identify influential observation in regression model. This approach starts with an initial subset of the data which are assumed to be free from outliers. Influential observations are identified by looking at the changes on parameter of interest as more observations are included into the subset. In this study, we extend the use of forward search method to identify influential observations in the survival regression problem. The initial subset and the order of observations entering the subset are determined by the order of squared residuals for every observation. The effects of observations entering the subset on the statistics of interest are observed through the progression plots. We suggest a simple statistic to identify observation which has large effect on the statistics of interest.

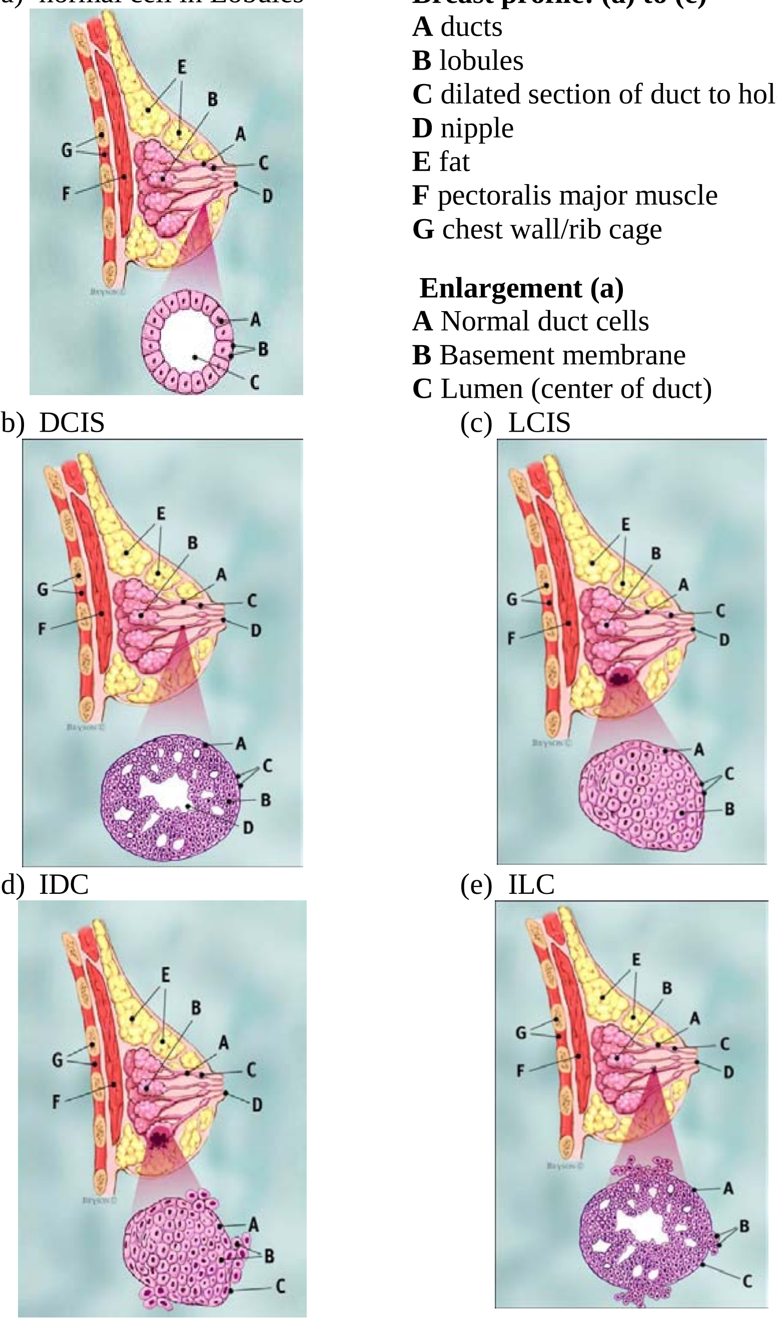
1.4 Breast Cancer

Cancer occurs as a result of abnormal changes in the genes responsible for regulating the growth of cells and keeping them healthy. The genes are in each cell’s nucleus, which acts as the “control room” of each cell. Normally, the cells in our bodies replace themselves through an orderly process of cell growth where healthy new cells take over as old ones die out. But over time, abnormal changes can “turn on” certain genes and “turn off” others in a cell. These cells gains ability to keep dividing without control or order, producing more cells just like it and forming a tumour (Weiss (2000)).

A tumour can be benign (not dangerous to health) or malignant (has the potential to be dangerous). Benign tumours are not considered as cancer because the cells are close to normal cells in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumours are cancerous and it can eventually spread beyond the original tumour to other parts of the body. The term “breast cancer” refers to a malignant tumour that has developed from cells in the breast and generally known as carcinoma in situ (CIS). Usually breast cancer either begins in the cells of the lobules (L), which are the milk-producing glands, or the ducts (D), the passages that drain milk from the lobules to the nipple. The malignant tumour can be non-invasive or invasive tumour. Invasive tumour is a tumour that has spread outside the milk duct or milk-making glands and has grown into normal tissue inside the breast (IDC or ILC) while, non-invasive tumour (DCIS or LCIS) has not spread beyond the milk duct into any normal surrounding breast tissue. The DCIS, LCIS, IDC and ILC images are as given in Figure 1.1.

Over time, cancer cells can invade nearby healthy breast tissue and make their way into the underarm lymph nodes, small organs that filter out foreign substances in the body. If cancer cells get into the lymph nodes, they then have a pathway into other parts of the body. The breast cancer stage refers to how far the cancer cells have spread beyond the original tumour. The situation of spreading of breast cancer will be discussed in detail in Chapter 4. To prevent cancer, there are steps that every person can take to help the body stay as healthy as possible such as seeing medical attention in early stages, eating a balanced diet, not smoking or exercising regularly.

Breast cancer is the most common cancer in women in most parts of the world. The World Health Organization predicts that more than 1.2 million people will be diagnosed with breast cancer worldwide in 2006. In 2000, there were 1,050,346 cases reported with 372,969 deaths from breast cancer world-wide. Variations in incidence have been attributed to variations in body size, diet and reproductive characteristics of women. In the second report of the National Cancer Registry which describe the occurrence of cancer among Malaysian in 2003, a total of 21,464 cancer cases were diagnosed in peninsular Malaysia. The crude incidence rate for females is 127.6 per 100,000 population with the standardized incidence rate at 154.2 per 100,000 females. The report had identified breast cancer as the most frequent cancer in females irrespective of ethnic group and age (more than 14 years old) with 3,738 new cases in 2003 alone (Lim *et. al* (2008)).



(a) normal cell in Lobules

Breast profile: (a) to (e)
A ducts
B lobules
C dilated section of duct to hold milk
D nipple
E fat
F pectoralis major muscle
G chest wall/rib cage

Enlargement (a)
A Normal duct cells
B Basement membrane
C Lumen (center of duct)

(b) DCIS

(c) LCIS

(d) IDC

(e) ILC

Enlargement (b)
A normal duct cells
B ductal cancer cells
C basement membrane
D lumen (center of duct)

Enlargement (c)
A normal lobular cells
B lobular cancer cells
C basement membrane

Enlargement (d)
A normal duct cells
B ductal cancer cells breaking through the basement membrane
C basement membrane

Enlargement (e)
A normal lobular cells
B lobular cancer cells breaking through the basement membrane
C basement membrane

Figure 1.1
Breast structure with cancer
Source: Weiss (2000)

In this study, we attempt to identify important prognostic factors that affect the survival of breast cancer patients in Malaysia. The prognostic factors are age, race, and pathological characteristics of the tumour. The identification of important prognostic factors can be done through modeling the survival times of patients (see Haybittle *et al.* (1982), Collet (1994), Lim *et al.* (2001) and Habibi *et al.* (2008)). The widely used model is the Cox PHM.

1.5 Problem Statement

Relatively little has been done on the identification of outliers and influential observations in survival studies. Unusual observation (outliers and influential observations) may have impact on the analysis. This study looks at different types of residuals that can be used to identify outliers in Cox PHM. To identify influential observations in the model, we propose to employ the forward search method which has been successfully used for similar purposes in other type of regression modeling.

1.6 Research Objectives

This study has three objectives:

1. To compare the procedures of detecting outliers based on different types of residuals for Cox PHM.
2. To propose a procedure in identifying influential observation using the forward search method.
3. To illustrate the application of the proposed procedure on local breast cancer data.

1.7 Thesis Outline

The outline of this study is:

Chapter 2 presents a review on survival data modeling, and outlier and influential observation detection in the Cox PHM. Special meanings of outliers in survival data are presented.

Chapter 3 presents the theory of Cox regression model including the parameter estimation, the baseline hazard estimation and the model selection.

Chapter 4 presents the survival analysis of Malaysia breast cancer data which are obtained from University of Malaya Medical Centre (UMMC).

Chapter 5 presents three types of residuals used to detect outliers and the delete-case method used in detecting influential observations in the Cox PHM. In addition, simulation study is conducted to see the sampling behavior of distributions of three types of residuals in the present of censoring. Further, we study the sampling behaviour of the minimum/maximum values of the residuals which can then be used as cut points for detecting outliers in the Cox PHM. We then use the prostate cancer data as an illustration.

Chapter 6 reviews the forward search method that has been used in detecting influential observations in generalized linear model. This method is then extended to the Cox PHM for similar purpose.

Chapter 7 presents the application of the outlier and influential observation detection procedures on two cohorts of local breast cancer data.

Chapter 8 concludes the report of the study with a summary of the research, contribution and future research.