CHAPTER TWO

LITERATURE REVIEW

2.1 Modeling Survival Data

Medical statisticians and actuarial scientists have described extensively the mortality experience of population using life-tables several centuries ago. Kaplan and Meier (1958) gave a comprehensive review of earlier work. They considered the problem of estimating a distribution function from censored data in mathematical statistics. Beside, Cox (1972) proposed a model that incorporates regression-like arguments for modeling survival data. The application of the model can be found in industrial reliability study, medical statistics and actuarial science. It is common to focus the attention on the effect of explanatory variables on survival experience of the individual. This model is known as Cox PHM as briefly described in Chapter 1.

The Cox PHM is given by $h(t; \boldsymbol{x}_i) = \exp(\boldsymbol{x}_i\boldsymbol{\beta})h_0(t)$ where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients of the model to be estimated from the conditional likelihood, $\boldsymbol{x}_i$ is a matrix of covariates for the ith individual and $h_0(t)$ is the baseline hazard function. Note that, $\exp(\boldsymbol{x}_i\boldsymbol{\beta})$ is a linear component of the model and is greater than zero. This model is also known as a semi-parametric model. The relevant likelihood function of the proposed model in estimating the parameter of Cox PHM is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{r} \left[\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{(i)})}{\sum_{l \in R(t)} \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)} \right] \tag{2.1}$$

where $\boldsymbol{x}_{(i)}$ is the vector of covariates for the individual who dies at the ith ordered death time $t_{(i)}$ and $R(t)$ is the set of individuals who are at risk time $t_{(i)}$. Here the product is over all uncensored failure time t_i and each term represents the conditional probability that the ith individual fails at time t_i given that one individual among those at risk at time t_i fails at time t_i. Equation (2.1) has been suggested to be treated as an ordinary likelihood function for the large-sample inference about $\boldsymbol{\beta}$ (see Cox (1975), Tsiatis (1981), Næs (1982), Bailey (1983) and Pons (2003)). Kalbfleisch and Prentice (1973) extended the work on the Cox PHM to estimate the regression parameters using a marginal likelihood approach in case when ties occur in the set of survival data.

Martinussen (1999) proposed the EM-algorithm to estimate the parameter of Cox PHM with missing values in the covariates. He specified a full model by letting the unobserved covariate values be random and maximized the observed likelihood. The asymptotic covariate is estimated by the inverse information matrix. The missing data are allowed to be missing at random but the non-ignorable non-response situation may also be considered in principle.

Andersen and Gill (1982) extended the Cox's idea to the covariate processes that have a proportional effect on the intensity process and multivariate counting process. This permits a statistical regression analysis of the intensity of a recurrence event allowing for complicated censoring patterns and time dependent covariates. This formulation gives rise to proofs with very simple structure using martingale techniques for the asymptotic properties of the estimators.

2.2 Proportional Hazard Assumption (PHA)

Schoenfeld (1982) proposed the partial residuals for Cox regression Model in investigating the proportional hazard assumption and also for examining and detecting outlying covariate values.

Andersen (1991) stated that there are two basic assumptions in modeling survival data using Cox proportional hazard data. Firstly, the effect of covariate $\boldsymbol{X}$ is linear and, secondly, $\boldsymbol{\beta}$ is constant over time. The latter constitutes the actual proportional hazard assumptions. Less explicitly, it is also assumed in Cox PHM that individuals are independent and homogeneous given their covariates.

Breslow *et al.* (1984) suggested a PHA test based on the Cox model with time-dependent covariate, where the probability that an event (death) comes from a certain group is calculated. These probabilities are weighted with either ranks of times or with cumulative hazard. A non-significant result on the time-dependent variable suggests that PHA is met.

Harell (1986) developed a test of proportional hazard assumption based on Schoenfeld partial residuals. The test used the Fisher's z-transform of Pearson correlation between the partial residuals and the rank order of failure times. The null hypothesis is that the correlation equals 0. On the other hand, Grambsch and Therneau (1994) developed a PHA test based on similar type of residual by measuring the difference between the observed and expected value of the covariate at each time of a standard time-independent Cox regression model. The residuals are then weighted as a time-function.

Gill and Schumacher (1987) proposed a test of the proportional hazard assumption based on a comparison of different generalized rank estimators of the relative risk. The null hypothesis of proportional hazard is H_o: $h_i(t,\mathbf{X^*}) / h_i(t,\mathbf{X}) = \theta$ for some positive constant θ, where $\boldsymbol{X^*}$ and $\boldsymbol{X}$ are the covariates of two different groups. Two different weight functions are used to estimate θ. The difference is then calculated and simply tested if it differs from 0.

Meanwhile, Quantin *et al.* (1996) proposed a global test of the proportional hazard assumption based on semi-parametric generalization of the proportional hazard regression model. The hypothesis is tested by using score statistic derived from the partial likelihood. Later Verweij *et al.* (1998) proposed the goodness-of-fit test on basis of martingale residuals for the Cox PHM. The test is derived as the score test for the presence of an extra random effect in the exponential part of the hazard function.

In general, a visual assessment of the validity of the PHA can be made from several types of plot as follows:

1) A plot of survival curves based on the Cox model and Kaplan- Meier estimates for each group.
2) A plot of cumulative baseline hazard in different groups and also known as the Andersen plot (see Andersen (1982) and Marubini and Valsecchi (1995)).
3) A plot of the difference of the log cumulative baseline hazard versus time.
4) A smoothed plot of the ratio of log cumulative baseline hazard rates versus time.
5) A smoothed plot of scaled Schoenfeld residuals versus time.
6) A plot of estimated cumulative hazard versus number of failures (see Arjas (1988)).

2.3 Outliers and influential observations

Hawkins (1980) defined an outlier as an observation that deviates so much from other observations as to arouse doubt that it was generated by a different mechanism. Outlier detection algorithms often fall into one of the categories of distance-based methods, density-based methods, projection-based methods, and distribution-based methods. A general approach to identify outliers is to assume a known distribution for the data and to examine the deviation of individuals from the distribution. Early work on outliers was carried out from a statistical viewpoint where outliers may deviate significantly from the identified underlying distribution of the data (see Beckman and Cook (1983) and Barnett and Lewis (1984)).

Therneau *et al.* (1990) proposed deviance residual for survival model. Deviance residual can be obtained from the martingale residual. It is given by

$$r_{Mi} = \delta_i - \left(\exp\left\{\hat{\boldsymbol{\beta}}' x_i\right\}\right)\hat{H}_o(t_i)$$

where the δ_i is the status of the *i*th individual (1 if died or 0 if censored) and the $\hat{H}_o(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i and $i = 1,2,...,n$. It is known that the martingale residuals are not symmetrically distributed resulting in the difficulty to screen or detect individuals that are poorly predicted by the model. On the other hand, the deviance residual is more symmetrically distributed and very close to a normal distribution. The deviance residuals is defined as

$$d_i = sign\left(\hat{M}_i\right)\left[-2\left\{\hat{M}_i + \delta_i \log\left(\delta_i - \hat{M}_i\right)\right\}\right]^{\frac{1}{2}}$$

where $\hat{M}_i$ is the estimated value of martingale residual for the *i*th individual. Therneau *et al.* (1990) further pointed out that when censoring in simulated data is less than 25%,

deviance residuals is very close to a normal distribution, but when censoring is greater than 40% there are large points with residuals near 0 that distorts the normal approximation. The same residual is further used by Therneau *et al.* (1990) and Therneau and Grambsch (2000) in detecting the outliers in Cox regression model and generalization of the Cox regression model setting for data analysis. Later, Nardi and Schemper (1999) proposed two types of residuals known as the log-odds residuals and normal deviate residuals for Cox PHM. The residuals are developed by transforming the survival probability using the logit and probit transformations. The residuals are then used to detect outliers in the Cox PHM (see Nardi and Schemper (1999) and Nardi and Schemper (2003)).

Several authors noted that the influential observations in survival analysis are often found among long survivors (see Therneau *et al.* (1990), Marubini and Valsecchi (1995) and Collet (2003)). The case deletion diagnostics is the standard procedure that is being used in identifying influential observations in generalized linear regression. It was introduced by Belsley *et al.* (1980). The procedure observes the changes in the estimated parameters of interest when an individual is dropped one at a time (see Hadi (1992) and Imon (2005)). Any significant changes when observation is dropped indicate that the said observation is an influential observation.

The same procedure can be extended to Cox PHM. The changes of the parameter coefficient $\Delta\hat{\beta}_{(j)}$ produced by dropping the *j*th observation in the fitted Cox PHM can be observed by plotting $\Delta\hat{\beta}_{(j)}$ versus j for $j = 1,2,...,n$. Meanwhile, Cain and Lange (1984) and Reid and Crépeau (1985) proposed the use of influence function in detecting influential observations in Cox PHM. The influence function diagnostic $\Delta\hat{\beta}_{(j)}^{IF}$ is obtained when the *j*th individual is dropped from the data. For visualization

purpose, they use the plot of $\Delta\hat{\beta}_{(j)}^{IF}$ versus j for $j = 1,2,...,n$. At the same time Storer and Crowley (1985) proposed the augmented (AUG) approach involving time-dependent covariates to identify influential observations in Cox PHM. The augmented diagnostic denoted by $\Delta_i\hat{\beta}_{(J)}^{AUG}$ is produced by eliminating the jth observation from the data. To see the effect, they use the plot of $\Delta_i\hat{\beta}_{(J)}^{AUG}$ versus j for $j = 1,2,...,n$. If any large changes are observed in any of the plots, the corresponding observation is identified as an influential observation. The performance of the three types of case-deletion methods were compared by Wang *et al.* (2006). Other approaches include added variable plot and variable plot (see Chen and Wang (1991)) and the pair wise deletion approach (see Wei and Kosorok (2000)).

2.4 Forward Search Method

Atkinson and Riani (2000) has suggested a very powerful method known as forward search method (FS method) on detecting influential observations in simple and multiple linear regression problems. Based on the residuals obtained from the fitted model, an initial subset of size m (m is smaller than the size of data set) and free from outliers is formed from the data. The effect of adding one observation at a time into the initial subset on the statistics of interest is continuously monitored until all observations are in the subset. Including influential observation is expected to cause some significant changes in the estimates of the statistics.

Atkinson and Riani (2001) proposed a simple robust method for detecting influential observations in binomial data. The technique is based on a forward search procedure which orders the observations from those most in agreement with a specified generalized linear model to those least in agreement with the model. The effectiveness

of the forward search estimator in detecting masked multiple outliers, and more generally in ordering binomial data, is shown by means of three data sets. Plots of diagnostic quantities during the forward search clearly show the effect of individual observations on residuals and test statistics. These examples reveal the strength of the method in describing the data where the FS method is simpler and more effective than the standard deletion diagnostic procedures. Atkinson and Riani (2002) also proposed FS on detecting masked outliers on model selection in the regression model. The method is able to highlight the effects of individual observations on the *t*-tests for the variables included in the model. The FS method has also been successfully used in comparing the effect of transforming data with outliers. Atkinson *et al.* (2004) had also used FS procedure in detecting masked outliers in multivariate regression model.

Besides, the FS method has been employed in several areas. For example, Crosilla *et al.* (2007) used FS method in searching outlier in spatial objects cases. Coin (2008) used FS method in testing normality assumption of univariate data to investigate the presence of outliers. In this study, we extend the application of FS method to identify the influential observation in Cox PHM.

2.5 Breast Cancer

It is known that breast cancer is the most frequent type of cancer in women both in the developed and the developing world. Many studies have been conducted to control this disease. For example, Haybittle *et al.* (1982) used Cox PHM on studying the effect of nine prognostic factors to primary operable breast cancer survival. Only three of the factors; tumour size, stage of disease, and tumour grade remained significant on the analysis. As a result, they proposed a prognostic index in discriminating patients into

three group of diagnosis. The index is then used in selecting suitable treatment for the primary tumour breast cancer patients. The model was further used by Galea *et al.* (1992), Ravichandran *et al.* (2005), Aryandono *et al.* (2006), Habibi *et al.* (2008) and Ford *et al.* (2008). Besides, there are cases when the Cox PHM is not appropriate in modeling breast cancer data. Therefore, researchers have also considered other survival model such as the stratified Cox PHM. They include Hatteville *et al.* (2002), Rosenberg *et al.* (2005) and Ejlertsen *et al.* (2007).

In Malaysia, research on breast cancer problem is still new. A research was conducted on studying the demographic pattern of local breast cancer and the factors of delaying presentation (see Hisham and Yip (2004)). The findings are used to promote the early detection of breast cancer, as well as in understanding the geographical and racial variations in the incidence of breast cancer. Yet there is no publication on modeling the performance of treatments on breast cancer survival or prevention study on breast cancer issues based on Cox PHM. Thus, this study attempts to model the local breast cancer data and consequently identifies significant prognostic factors that may affect the survival of breast cancer patients based on survival model particularly the Cox PHM. The analysis is presented in Chapter 4.