Appendix B
Estimation of Parameters and Baseline Hazard Function in Proportional Hazards Model

B.1 Estimation of the Coefficient $\underline{b}$

Among the proposed methods for estimating the parameters $\underline{b}$, those of marginal and partial likelihood are the most commonly used in practice. Since the partial likelihood method developed by Cox (1972) is considered to be the most general of the existing estimation techniques and will be used in the data analysis involved in the present study, further details on its derivation are given in this section.

Consider the following two scenarios. In scenario 1 an item fails at 100 hours. In scenario 2 an item fails at 100 hours given that it has survived for 99 hours. In the first scenario we use the failure time density to evaluate the corresponding probability function. In the second, we need to use the hazard function to evaluate the conditional probability.

Now let us consider the event that item i fails from the survivor or risk set R(t_j) and R(t_j) = {k: $t_k \geq t_j$} in which: t_k is the failure time for item k and $t_1 < t_2 < \ldots < t_n$ denotes the ordered failure times of the n pipes. To evaluate the corresponding probability, we first evaluate the probability that one item fails in the survivor set R(t_j) which is given by:

$$\text{P[one item fails at t from R}(t_j)] = \sum_{k} \text{P[item k fails at time } \tau_j \text{ from R}(\tau_j)]$$
$$= \sum_{k \in R(t)} P[\tau_j \leq t_k < \tau_j + \Delta t \,|\, t_k \geq \tau_j]$$
$$= \sum_{k \in R(\tau_j)} h_k(t)\Delta t$$

Consider the probability, P[item 'i' fails at time t_j given that it belongs to R(t_j) | one item fails in R(t_j)] in which R(t_j) ={k: $t_k \geq t_j$} which in turn can be evaluated as

$$\frac{\text{P[item } k \in R(\tau_j) \text{ fails]}}{\text{P[one item fails in R}(\tau_j)]}$$
$$= \frac{P[t_j \leq t_k < t_j + \Delta t]}{P[t_k \geq t]\,\text{P[one item fails in R}(t_j)]}$$
$$= \frac{h_i(\tau_j)}{\sum_{k \in R(\tau_j)} h_k(\tau_j)}$$

By regarding the selection of an item from the risk set at its observed failure time as an independent event, a joint probability statement more appropriately called the likelihood function is written as:

$$L(\underline{\beta};t_1,t_2,...,t_n)=\prod_{i=1}^{n}\frac{h_i(t)}{\sum_{k\in R(t_i)}h_k(t)}$$

By applying the proportional hazards model structure, $h(t;z) = h_0(t)y(\underline{z};\underline{b})$, the likelihood function becomes

$$L(\underline{\beta})=\prod_{j=1}^{n}\frac{h_0(\tau_j)\exp(\underline{\beta}\underline{z}_i)}{\sum_{k\in R(\tau_j)}h_0(\tau_j)\exp(\underline{\beta}\underline{z}_k)}=\prod_{j=1}^{n}\frac{\exp(\underline{\beta}\underline{z}_i)}{\sum_{k\in R(\tau_j)}\exp(\underline{\beta}\underline{z}_k)} \tag{B.1.1}$$

Several methods have also been proposed for dealing with the problem of ties in failure times [Cox and Oaks (1984), Kalbfleisch and Prentice (1980)]. In the approximate likelihood function in the case of ties as proposed by Breslow (1974), the partial likelihood function is written as:

$$L(\underline{\beta})=\prod_{i=1}^{n}\frac{\exp(\underline{\beta}\underline{s}_i)}{\left[\sum_{j\in R(\tau_j)}\exp(\underline{\beta}\underline{z}_j)\right]^{m_i}}$$

where m_i is the number of failures at t_i and s_i is the vector sum of the covariates of the m_i items.

The conditional log likelihood obtained from equation (B.1.1) is given by

$$L(\underline{\beta})=\sum_{i=1}^{n}\underline{\beta}\underline{z}_i-\sum_{i=1}^{n}\log\left[\sum_{k\in R(\tau_j)}\exp(\underline{\beta}\underline{z}_k)\right]=\sum_{i=1}^{n}l_i \tag{B.1.2}$$

In order to obtain maximum likelihood estimates of $\underline{b}$ the first and second derivatives of l_i are first calculated. If z_{ir} denotes the value of the rth component of the explanatory variable z on the ith subject, then we have:

$$\frac{\partial l_i}{\partial l_r}=z_{ir}-\frac{\sum_{k\in R(\tau_i)}z_{kr}\exp(\underline{\beta}^T\underline{z}_k)}{\sum_{k\in R(\tau_i)}\exp(\underline{\beta}^T\underline{z}_k)}=z_{ir}-A_{ir}(\underline{\beta}) \tag{B.1.3}$$

Also:

$$\frac{\partial^2 l_i}{\partial\beta_r\partial\beta_s}=z_{ir}-\frac{\sum_{k\in R(\tau_i)}z_{kr}\exp(\underline{\beta}^T\underline{z}_k)}{\sum_{k\in R(\tau_i)}\exp(\underline{\beta}^T\underline{z}_k)}+A_{ir}(\underline{\beta})A_{is}(\underline{\beta})=-C_{irs}(\underline{\beta})$$

For all risk sets i = 1, … , k, we obtain from equation (B.1.3) the score function $U_r(\underline{b})$ for the rth component:

$$U_r(\underline{\beta}) = \sum_{i=1}^{k} \frac{\partial l_i}{\partial \beta_r} = \sum_{i=1}^{k} \left(z_{ir} - A_{ir}(\underline{\beta}) \right) \tag{B.1.4}$$

The information matrix I($\underline{b}$) of negative second derivatives has elements:

$$I_{rs}(\underline{\beta}) = \sum_{i=1}^{k} C_{irs}(\underline{\beta}) \tag{B.1.5}$$

Maximum likelihood estimates of $\underline{b}$ can be obtained by iterative use of equation (B.1.4) and (B.1.5). The Newton-Raphson iterative algorithm can be applied to obtain the maximum likelihood estimates of $\underline{b}$.

B.2 Estimation of the Baseline Hazard Function

By referring to equation (5.1.5), it is seen that we still need to estimate the baseline hazard function $h_0(t)$. From equation (5.1.2), it is seen that the baseline hazard function represents the hazard rate that a piece of equipment would experience if all the values of the covariates are equal to zero. We can say that the baseline hazard function is the hazard rate when the covariates have no influence on the failure pattern.

The baseline hazard function can be estimated by first obtaining a cumulative baseline hazard function.

First, compute the cumulative hazard function as follows. From the given data we know the observed failure times, t_j . Then for a fixed value of t, $H_0(t)$ is estimated by

$$\hat{H}_0(t) = \sum_{\tau_j < t} \frac{d_j}{\sum_{k \in R(\tau_j)} \exp(\hat{\underline{\beta}} \cdot \underline{z}_k)} \tag{B.2.1}$$

where,

t_j : observed and ordered failure times

d_j : number of failures at time t_j

R(t_j) : risk set at time t_j = {k: $t_k \geq t_j$ }

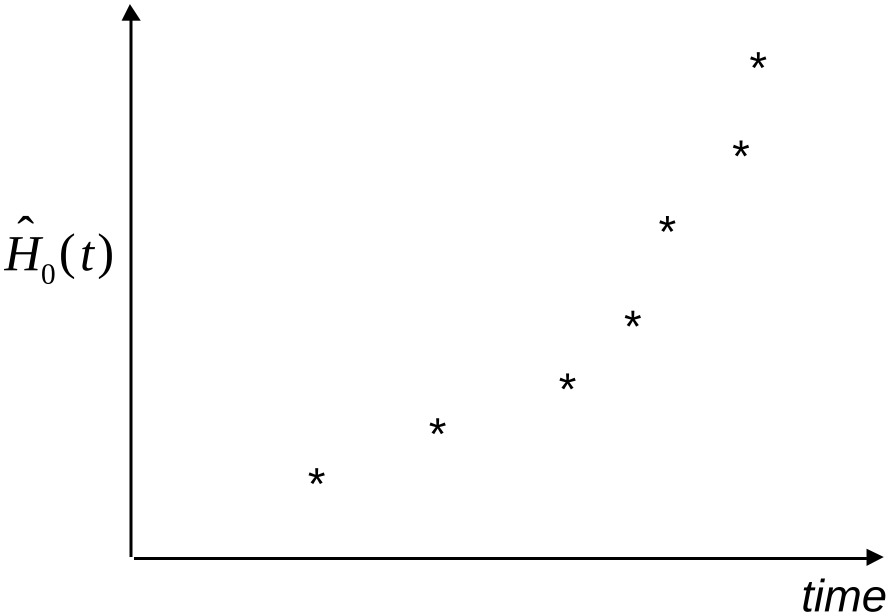

Figure B.1 Plot of time versus $\hat{H}_0(t)$

Second, plot time, t, versus computed values of $\hat{H}_0(t)$ at each failure time as shown in Figure B.1. A curve (equation) can be fitted through the discrete data points for a continuous $\hat{H}_0(t)$. Finally, we take the derivative of $\hat{H}_0(t)$ with respect to time, t, to obtain the estimate of the baseline hazard function:

$$\hat{h}_0(t) = \frac{d\hat{H}_0(t)}{dt}$$

The Proportional Hazard Model can be obtained by multiplying the estimated covariate structure $e^{\underline{\hat{b}z}}$ and $\hat{h}_0(t)$. Therefore, the estimated Proportional Hazards Model is expressed as

$$\hat{h}(t;\underline{z}) = \hat{\mathrm{h}}_0(t)\exp(\underline{z}\,;\underline{\hat{b}})$$

In summary, the coefficients of Proportional Hazards Model and the baseline hazard function are obtained by the following steps.

(1) The coefficients of covariates, $\underline{b}$, are obtained by maximizing the log likelihood function (eq. (B.1.2))
(2) The baseline hazard function is obtained by first obtaining the cumulative baseline hazard function by using eq. (B.2.1). By taking derivative of eq. (B.2.1) we obtain the baseline hazard function.