國 立 交 通 大 學

統計學研究所

碩 士 論 文

區間設限資料之統計推論-文獻回顧

Statistical Inference based on Interval Censored Data A Literature Review

研 究 生：黃祥福

指導教授：王維菁　教授

中 華 民 國 一 百 年 五 月

區間設限資料之統計推論–文獻回顧

Statistical Inference based on Interval Censored Data

A Literature Review

研　究　生：黃祥福　　Student： Hsiang-Fu Huang

指導教授：王維菁 教授　　Advisor：Dr. Wei-Jing Wang

國　立　交　通　大　學

統　計　學　研　究　所

碩　士　論　文

A Thesis

Submitted to Institute of Statistics

College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

May 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年五月

區間設限資料之統計推論

研究生：黃祥福　　指導教授：王維菁 教授

國立交通大學理學院

統計學研究所

摘要

在此論文中，我們回顧區間設限資料的推論問題，為呈現概念的建構原則，亦簡述其它類型的不完整資料. 論文分兩大部份，一部份為無母數估計，另一部份為迴歸分析. 我們回顧了兩類的無母數估計法，其中自我一致演算法可視為動差法的延伸. 另一方法為無母數最大概似估計法. 我們介紹三種較廣泛使用的迴歸模型： 包含比例風險模型，加速失敗模型和比例勝算模型. 推論的困難度在於模式存在未知函數，需要利用平滑的技巧處理之. 本論文以介紹點估計的概念為主，並未涵蓋如何由分佈理論推導信賴區間與統計檢定問題.

關鍵字:區間設限， 無母數

Statistical Inference based on Interval Censored Data

Student：Hsiang-Fu Huang Advisor：Dr. Wei-Jing Wang

Institute of Statistics
National Chiao Tung University
Hsinchu, Taiwan

Abstract

We review inference methods for analyzing incomplete data with focus on interval censored data. For nonparametric analysis, two estimation approaches are examined. Self-consistency can be viewed as an extension of the method of moment by imputing incomplete information by its expected value. The other is the nonparametric likelihood estimation. We also introduce three popular regression models, namely the proportional hazards model, accelerated failure time model, and proportional odds model. These models contain unknown nuisance functions and different smoothing techniques are employed to handle them in the estimation procedure. The thesis focuses on point estimation so that second ordered properties are not investigated.

Keywords: interval censoring; nonparametric

致　　謝

很榮幸能夠成為王維菁 教授的指導學生，在跟老師寫論文的這一年當中，老師總是不餘遺力的教導和督促我們．天性散漫的我受到老師的影響，態度也變得較積極．在課業方面，老師除了幫忙解惑之外，還訓練我如何去找問題的核心所在，進而去尋找解決問題的方法．這讓我從一味只會全盤接受他人的想法，慢慢的進步到去思考別人的想法的正確性，對我的獨立思考能力有相當的幫助．此外，表達能力也是老師教學的重點．這些都是我先前相當缺乏的能力．雖然我在碩二這一年沒有修課，但是有這些訓練，讓我覺得比在課堂上學的更多．相信在之後的職場上也有很大的幫助．

此外也要特別感謝李博文 學長和蘇健霖 學長對我論文的協助，我的論文才能在時間內順利完成．

黃祥福 僅誌

中華民國 一百年夏

于交通大學統計學研究所

Contents

Chapter 1. Introduction

1.1 Motivation

In survival analysis we are interested in studying the behavior of the time to the event of interest, denoted as T, which often cannot be completely observed. Textbooks on survival analysis focus on right censored data in which many elegant results have been derived. In the thesis, we consider interval censored data. Although such data are commonly seen in practical applications, related statistical methods are not formally taught in the classroom. The lack of an overall review at the introductory level provides the motivation of the thesis.

To enhance a thorough understanding about the methodology, we will still review related materials for complete and right censored data and address how the basic ideas are modified when the data become interval censored. We will consider two types of applications. One is nonparametric analysis of the survival function $\Pr(T > t)$. The other is regression analysis. Specifically let Z be the covariate vector which provides individual information. A regression model imposes additional assumptions on how Z affects T. We will review statistical methods for the proportional hazards (PH) model, the accelerated failure time (AFT) model and the proportional odds model when T is subject to interval censoring.

1.2 Different types of incomplete data

It is worthy to introduce different types of incomplete data and compare their differences.

(1). Right censored data

Observed variables can be expressed as (X, δ) where $X = T \wedge C$ and $\delta = I(T \le C)$. Thus when $\delta = 1$, $X = T$ and $X < C$; when $\delta = 0$, $X = C$ and $X < T$ (the true failure time T is located in the right-hand side of the observed time X). Right censoring occurs due to subject's withdrawal or the end-of-study effect.

Most literature in survival analysis consider right censored data.

(2) Left censored data:

We use the same notations to denote observed variables which have different definitions. One observes (X,δ) where $X = T \vee C$ and $\delta = I(T \geq C)$. When $\delta = 1$, $X = T$ and $X > C$; when $\delta = 0$, $X = C$ and $X > T$ (the true failure time T is located in the left-hand side of the observed time X). Here is an example of left censoring. Denote T as the time until first marijuana use among high school boys in California. If the event had already occurred prior to the recruitment of the study, the subject is said to be left censored.

(3) Doubly censored data

Doubly censored data include three types of observations, namely "observed", "right censored" and "left censored" ones. Let C_r be the right censoring variable and C_l be the left censoring variable. It is assumed that $C_l < C_r$. Observed variables are (X,δ) where $X = \max(\min(C_r, T), C_l)$ and $\delta = 1$ if $X = T$; $\delta = 0$ if $X = C_r$; and $\delta = -1$ if $X = C_l$.

(4) Interval censored data

The observation is a random interval $(L,R]$ such that T falls in the interval but its exact value is unknown unless $L = R$. Such data can include right censoring if $R = \infty$ and left censoring if $L = 0$. Interval censored are commonly seen in practice. For example a subject is under study has a regular appointment schedule (say every three months) to the hospital. The left endpoint L refers to the last follow-up time that the event has not occurred and the right endpoint R refers to the first follow-up time that the event is detected.

1.3 Parametric analysis for interval censored data

The thesis focuses on nonparametric and semi-parametric methods. However these methods are still motivated by the parametric approach. Here we briefly summarize the parametric analysis for interval censored data denoted as $\{(L_i, R_i]\ (i = 1,...,n)\}$. Assume that the distribution function of T is specified as $F_\theta(\cdot)$. The likelihood function and the log-likelihood function can be written as

$$L(\theta) = \prod_{i=1}^{n} [F_\theta(R_i) - F_\theta(L_i)] \qquad (1.1$$

and

$$\log L(\theta) = \sum_{i=1}^{n} \log\{F_\theta(R_i) - F_\theta(L_i)\}. \qquad (1.2)$$

The resulting score function is

$$S(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{f_\theta(R_i) - f_\theta(L_i)}{F_\theta(R_i) - F_\theta(L_i)}. \qquad (1.3)$$

By solving $S(\theta) = 0$, we obtain the maximum likelihood estimator of θ. Our analysis implies that parametric inference is straightforward for interval censored data. In fact, difficulties arise when the functional form of $F(\cdot)$ is completely unspecified as in the nonparametric setting or partially specified as in the semi-parametric setting.

1.4 Organization of the Thesis

Chapter 2 considers non-parametric inference and Chapter 3 to Chapter 5 discuss semi-parametric regression models, namely the proportional hazards (PH) model, the accelerated failure time (AFT) model and the proportional odds (PO) model respectively. Chapter 6 gives a brief summary of the thesis.

Chapter 2. Inference based on Nonparametric Analysis

The major objective is to estimate the survival function $S(t) = \Pr(T > t)$ without specifying the parametric distribution. We will review nonparametric methods based on several data structures, from the simplest to the complicated data settings. This will help us to understand the fundamental techniques more thoroughly.

2.1 Complete data

Complete data are denoted as $\{T_i \ (i = 1, ..., n)\}$ which form a random sample of T. Based on the fact that $S(t) = E[I(T > t)]$ and applying the method of moment, one can estimate it by the empirical estimator

$$\bar{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t).$$

Although the method of moment provides a simple solution, it does not guarantee any optimality property. Formally the nonparametric likelihood estimator should be pursued as a better option. The first step is to re-write the likelihood function based on grid points and the probabilities at these points are the parameters to be estimated. Suppose that there are only M distinct observed values denoted as $t_{(1)} < ... < t_{(M)}$. Let $d_j = \sum_{i=1}^{n} I(T_i = t_{(j)})$ be the number of observations at $t_{(j)}$ and $p_j = S(t_{(j)}-) - S(t_{(j)})$. The likelihood function can be written as

$$L = \prod_{j=1}^{n} \Pr(T = t_{(j)}) = (p_1)^{d_1}(p_2)^{d_2}...(p_M)^{d_M}. \qquad (2.1)$$

The NPMLE is the solution which maximizes $L(p_1, ..., p_M)$ subject to $\sum_{j=1}^{M} p_j = 1$. Applying the theorem of Lagrange multipliers, let $g(p_1, ..., p_M) = \log L - \lambda(\sum_{i=1}^{M} p_i - 1)$ and set $\nabla g(p_1, ..., p_M) = 0$ Hence the Lagrange equation system is given by

$$\frac{d_j}{p_j} - \lambda = 0 \ (j = 1, ..., M) \text{ and } \sum_{j=1}^{M} p_j = 1. \qquad (2.2$$

The above equations can be solved easily and the solution is $\hat{p}_j = d_j / n$ for $j = 1, ..., M$ and $\lambda = n$.

2.2 Right censored data

For right censored data, observed data become $\{(X_i, \delta_i)\ (i = 1, ..., n)\}$, where $X_i = \min(C_i, T_i)$ $\delta_i = I(T_i \le C_i)$. The survival function can be estimated by the Kaplan-Meier estimator

$$\hat{S}(t) = \prod_{u \le t} \left\{ 1 - \frac{\sum_i I(X_i = u, \delta_i = 1)}{\sum_i I(X_i \ge u)} \right\}.$$

The above product-limit expression is elegant since the censoring effect is cancelled out in estimation of the hazard function. However under other types of censoring mechanism, estimation of the hazards does not provide any advantage. Instead, the techniques of self-consistency and nonparametric MLE can be generalized to other data structures. Now we illustrate the self-consistency algorithm for right censored data. The self-consistency equation can be written as

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \left(I(X_i > t) + I(X_i \le t, \delta_i = 0) \frac{\hat{S}(t)}{\hat{S}(X_i)} \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(I(X_i > t) + (1 - \delta_i) I(X_i \le t) w_i \right). \qquad (2.3)$$

Notice that for points with $X_i > t$, the full weight 1 is assigned; for points with $X_i < t$ and $\delta_i = 1$, zero weight is assigned; and for points with $X_i < t$ and $\delta_i = 0$, partial weight $w_i = \hat{S}(t) / \hat{S}(X_i)$ is assigned. The estimator can be solved successively and explicitly. Since $\hat{S}(X_i) = 1$ for $X_i \le t_{(1)}$ with $\delta_i = 0$, we have

$$\hat{S}(t_{(1)}) = \frac{1}{n} \sum_{i=1}^{n} \left(I(X_i > t_{(1)}) + (1 - \delta_i) I(X_i \le t_{(1)}) \hat{S}(t_{(1)}) \right)$$

which allows us to solve $\hat{S}(t_{(1)})$ and then successively for $\hat{S}(t_{(j)})$ $j = 2,...,M$.

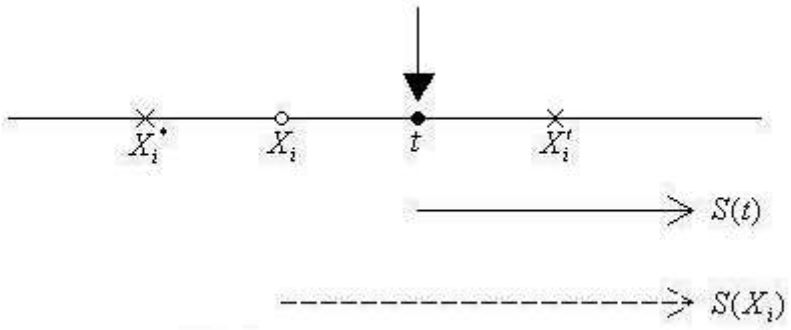


Figure 2.1 Self-consistency for right censored data.

Weight for $X_i' = 1$; weight for $X_i = S(t)/S(X_i)$; weight for $X_i^* = 0$

Now we discuss the nonparametric MLE approach for right censored data. Define $t_{(1)} < ... < t_{(M)}$ as distinct observed failure times and $\{x_{(j_1)},...,x_{(j_{m_j})}\}$ as ordered censored points in the interval $[t_{(j)}, t_{(j+1)})$. Also let $d_j = \sum_{i=1}^{n} I(X_i = t_{(j)}, \delta_i = 1)$ and $m_j = \sum_{i=1}^{n} I(X_i \in [t_{(j)}, t_{(j+1)}), \delta_i = 0))$. The likelihood function can be written as

$$L = \prod_{j=1}^{M} \left[\Pr(T = t_{(j)}) \right]^{d_j} \prod_{l=1}^{m_j} S(x_{(jl)}). \qquad (2.4$$

Since the survival function $S(t)$ is an non-increasing function, we can make the likelihood function larger when $S(x_{(j_l)})$ is replaced by $S(t_{(j)})$,This suggest that we can instead consider

$$L^* = \prod_{j=1}^{M} \left[S(t_{(j-1)}) - S(t_{(j)}) \right]^{d_j} \left[S(t_{(j)}) \right]^{m_j}.$$

Direct maximization of L^* in terms of $\Pr(t_{(j-1)} < T \le t_{(j)})$ or $S(t_{(j)})$ is not suggested for right censored data. Alternatively L^* can be re-parameterized in terms of hazards. Specifically we have $S(t_{(j)}) = \prod_{i=1}^{j}[1-\lambda(t_{(i)})]$ and hence

$$S(t_{(j-1)}) - S(t_{(j)}) = \prod_{i=1}^{j-1}[1-\lambda(t_{(i)})][\lambda(t_{(j)})].$$

It follows that

$$L^* = \prod_{j=1}^{M}\{\prod_{i=1}^{j-1}[1-\lambda(t_{(i)})][\lambda(t_{(j)})]\}^{d_j}\{\prod_{i=1}^{j}[1-\lambda(t_{(i)})]\}^{m_j}$$

$$= \prod_{j=1}^{M}[\lambda(t_{(j)})^{d_j}][\text{illegible}] \qquad (2.5)$$

where $n_j = \sum_{i=1}^{n} I(X_i \ge t_{(j)})$. Maximization in terms of $\lambda(t_{(j)}) = \lambda_j$, we obtain $\hat{\lambda}_j = d_j / n_j$ and accordingly

$$\hat{S}(t) = \prod_{t_{(j)} \le t}\{1-\hat{\lambda}(t_{(j)})\} = \prod_{t_{(j)} \le t}\{1-\frac{d_j}{n_j}\}$$

which is the Kaplan-Meier estimator.

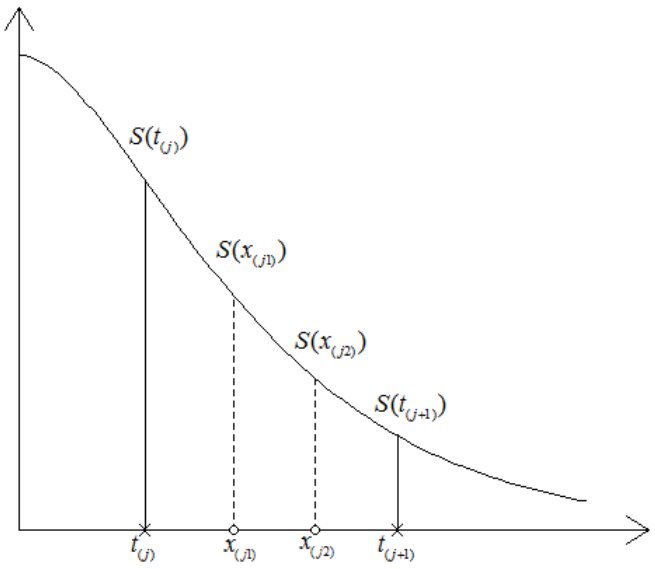


Figure 2.2:Idea for modifying the likelihood L

2.3 Double censored data

Recall that observed data can be denoted as: $\{(X_i,\delta_i)\ (i=1,...,n)\}$, where $X_i=\max\{\min(C_{ri},T_i),C_{li}\}$ and $\delta_i=-1$ if $X_i=C_{li}$; $\delta_i=1$ if $X_i=T_i$ and $\delta_i=0$ if $X_i=C_{ri}$. The self-consistency equation can be written as

$$\hat{S}(t)=\frac{1}{n}\sum_{i=1}^{n}I(X_i>t,\delta_i=1)+I(X_i>t,\delta_i=-1)\frac{\hat{S}(t)-\hat{S}(X_i)}{1-\hat{S}(X_i)}$$

$$+I(X_i\le t,\delta_i=0)\frac{\hat{S}(t)}{\hat{S}(X_i)}. \qquad (2.6$$

Turnbull (1974) suggested the following steps to implement the algorithm. First, partition the observation period: $(0,t_{(1)}]$, …, $(t_{(M-1)},t_{(M)}]$, where $t_{(1)}<...<t_{(M)}$ are ordered time points. Then define three types of observations in each interval

$$N_j^{(1)}=\sum_{i=1}^{n}I(X_i\in(t_{(j-1)},t_{(j)}],\Delta_i=1),$$

$$N_j^{(-1)}=\sum_{i=1}^{n}I(X_i\in(t_{(j-1)},t_{(j)}],\Delta_i=-1)$$

$$N_j^{(0)}=\sum_{i=1}^{n}I(X_i\in(t_{(j-1)},t_{(j)}],\Delta_i=0).$$

Turnbull's idea is to combine "exact" and "left censored" observations by imputation. Specifically for a point in $(t_{(k-1)},t_{(k)}]$ with $\Delta=-1$, the conditional probability of falling in $(t_{(j-1)},t_{(j)}]$ (for all $k\ge j$) is

$$\frac{F(t_{(j)})-F(t_{(j-1)})}{F(t_{(k)})}=\frac{S(t_{(j-1)})-S(t_{(j)})}{1-S(t_{(k)})}=\alpha_{kj}.$$

Thus the imputed value $\sum_{i=1}^{n}I(T_i\in(t_{(j-1)},t_{(j)}])$ is given by

$$\tilde{N}_j^{(1)}=N_j^{(1)}+\sum_{k=j}^{k=m}N_k^{(-1)}\hat{\alpha}_{kj}.$$

which is a function of data and $\hat{S}(.)$ Now the modified data become

$(\tilde{N}_j^{(1)}, N_j^{(0)})(j=1,...,m)$ which contains "pseudo-exact" and "right censored" points. Turnbull (1974) suggested to use them to estimate the hazard probability in the interval $(t_{(j-1)}, t_{(j)}]$ by $\tilde{N}_j^{(1)}/\tilde{Y}_j$ where $\tilde{Y}_j = \sum_{k=j}^{k=m} \tilde{N}_k^{(1)} + N_k^{(0)}$ denotes the number at risk and $\tilde{N}_j^{(1)}$ denotes the number of failure. The product-limit expression gives a direct relationship between the survival function and the hazard probabilities:

$$\hat{S}(t_{(j)}) = \prod_{i \le j} \{1 - \tilde{N}_i^{(1)}/\tilde{Y}_i\} \qquad (2.7$$

Note that since both $\tilde{N}_j^{(1)}$ and $\tilde{Y}_j$ are still functions of $\hat{S}(t_{(1)}),..., \hat{S}(t_{(M)})$, iterations are needed which can be summarized by the following steps.

Step 1: Give initial values: $\hat{S}^{(0)}(t_{(1)}),..., \hat{S}^{(0)}(t_{(M)})$;

Step 2: compute : $\hat{\alpha}_{kj}^{(0)}$

Step 3:obtain : $\hat{S}^{(1)}(t_{(1)}),..., \hat{S}^{(1)}(t_{(M)})$;

Step 4: repeat (2) and (3) until the pre-specified convergence criteria is reached.

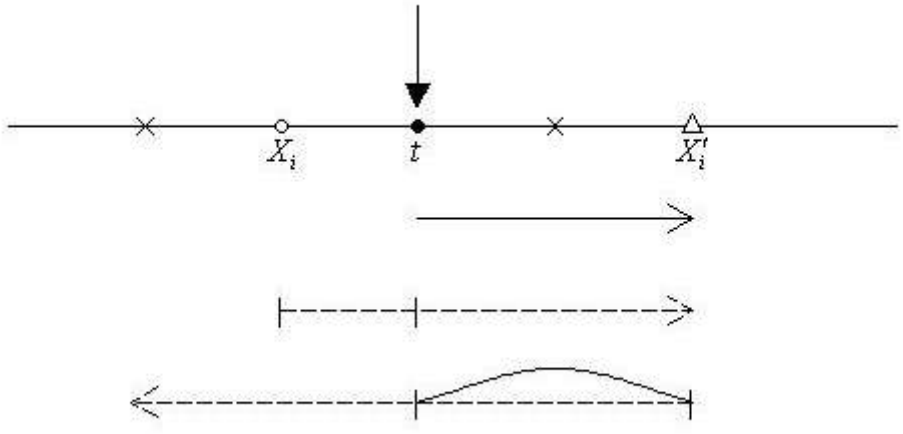


Figure 2.3 Self –consistency for double censored data

Weight for X_i = $S(X_i)/S(t)$; weight for $X_i' = \Pr(t < T \le X_i')/F(X_i')$

We discuss the NPMLE approach for doubly censored data. The following result is summarized form the work of Mykland and Ren (1996). Let the log-likelihood function of complete data be

$$l_c = \sum_{i=1}^{n} \log P(T = t_i) \qquad (2.8)$$

When the failure time $(T_1,...,T_n)$ is known, the NPMLE is easy to obtain as in section 2.1. But under doubly censoring, we cannot observe all the failure time. Hence the EM algorithm is applied. To obtain the iteration, we need to calculate

$$Q(P \mid P_0) = \sum_{i=1}^{n} E_{P_0}[\log P(T = t_i) \mid X_{(i)}, \delta_i]$$

where P_0 is the initial distribution of T. Here we do not group the observed data as in Turnbull(1974). Instead, we assume that the distribution of T has mass only at each observations $X_{(1)} \le \cdots \le X_{(n)}$. Then we discuss the following conditions.

(1) $E_{P_0}[\log P(T = t_i) \mid X_{(i)}, \delta_{(i)} = 1] = \log P_i(T = X_{(i)})$

(2) $E_{P_0}[\log P(T = t_i) \mid X_{(i)}, \delta_{(i)} = 0] = \dfrac{\sum_{X_{(i)} < X_{(j)}} P_j^{(0)} \log P(T = X_{(j)})}{S_i^{(0)}}$

(3) $E_{P_0}[\log P(T = t_i) \mid X_{(i)}, \delta_{(i)} = -1] = \dfrac{\sum_{X_{(i)} > X_{(j)}} P_j^{(0)} \log P(T = X_{(j)})}{1 - S_i^{(0)}}$

where $P_i^{(0)}$ and $S_i^{(0)}$ is the initial probability and survival at $X_{(i)}$, respectively. Hence for all distinct points $w_1 < w_2 < \cdots < w_M$ of the set $\{X_{(1)} \le \cdots \le X_{(n)}\}$,

$$Q(P \mid P_0) = \sum_{i=1}^{n} \left\{ I(\delta_i = 1) \log P_i(T = X_{(i)}) + I(\delta_i = 0) \frac{\sum_{i<j} P_j^{(0)} \log P(T = X_{(j)})}{S_i^{(0)}} \right.$$

$$\left. + I(\delta_i = -1) \frac{\sum_{i>j} P_j^{(0)} \log P(T = X_{(j)})}{1 - S_i^{(0)}} \right\}$$

$$= \sum_{j=1}^{n} \left\{ I(\delta_i = 1) + P_j^{(0)} \sum_{i=1}^{n} \frac{I(\delta_i = 0, X_{(i)} < X_{(j)})}{S_i^{(0)}} \right.$$

$$\left. + P_j^{(0)} \sum_{i=1}^{n} \frac{I(\delta_i = -1, X_{(i)} > X_{(j)})}{S_i^{(0)}} \right\} \log P(T = X_{(j)})$$

$$= \sum_{k=1}^{M} \left\{ \alpha_k + \pi_k P_k^{(0)} \sum_{i=1}^{n} \frac{I(\delta_i = 0, X_{(i)} < w_k)}{S_i^{(0)}} \right.$$

$$\left. + \pi_k P_k^{(0)} \sum_{i=1}^{n} \frac{I(\delta_i = -1, X_{(i)} > w_k)}{S_i^{(0)}} \right\} \log P(T = w_k)$$

where $\alpha_k = \sum_{i=1}^{n} I(X_{(i)} = w_k, \delta_i = 1)$, $\pi_k = \sum_{i=1}^{n} I(X_{(i)} = w_k)$ and $P_k^{\,0} = P^0(T = w_k)$. For simply the exception $Q(P \mid P_0)$ we define

$$\lambda_k = \alpha_k + \pi_k P_k^{(0)} \sum_{i=1}^{n} \frac{I(\delta_i = 0, X_{(i)} < w_k)}{S_i^{(0)}} + \pi_k P_k^{(0)} \sum_{i=1}^{n} \frac{I(\delta_i = -1, X_{(i)} > X_{(j)})}{S_i^{(0)}}$$

and

$$P(T = w_k) = q_k \quad k = 1, \ldots, M$$

Hence $Q(P \mid P_0)$ can be re-expressed as $\sum_{k=1}^{M} \lambda_k \log q_k$. Then the M-step is given by

$$\max \sum_{k=1}^{M} \lambda_k \log q_k \quad \text{subject to} \quad \sum_{k=1}^{M} q_k = 1 \tag{2.9}$$

Applying the theorem of Lagrange multiplier, the maximization is attained at

$$\hat{q}_k = \lambda_k / n \quad k = 1, \ldots, M .$$

The steps of iterations are summarized below:

Step 1: set $\hat{q}_k^{(0)} = 1/M \quad (k = 1, \ldots, M)$;

Step 2: set $\hat{q}_k^{(i)} = \lambda_k / n \quad (k = 1, \ldots, M)$;

Step 3: repeat $i = 2, 3, \ldots$ (Step 2) for until convergence.

2.4 Interval censored data

Recall that we observe $\{(L_i, R_i]\ (i = 1,...,n)\}$ and know that $T_i \in (L_i, R_i]$. Turnbull (1976) discussed nonparametric estimation of $\Pr(T > t)$ based on interval censored data. The first crucial step is to construct ordered disjoint intervals in which the probability can be estimated. To achieve this goal, the data $\{(L_i, R_i](i = 1,...,n)\}$ are re-arranged in an ascending order and the indentify the intervals such that a left-endpoint and a right-endpoint are adjacent to each other. Denote the disjoint intervals as $\{(l_j, r_j]\ (j = 1,...,m)\}$. It can be shown that only these intervals can receive positive mass. Note that an observation $(L_i, R_i]$ can occupy more than one intervals $(l_j, r_j]$ for some j. Denote the $\delta_{ij} = \mathbf{I}\{(l_j, r_j] \subset (L_i, R_i)\}$ which indicates whether $(L_i, R_i]$ overlaps with $(l_j, r_j]$. Denote $p_j = F(r_j) - F(l_j)$ as the mass in $(l_j, r_j]$.

The self-consistency equation for an estimator of p_j satisfies

$$p_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{ij} p_j}{\delta_{i1} p_1 + ... + \delta_{im} p_m} = \frac{1}{n}\sum_{i=1}^{n} w_{ij} \quad (j = 1,...,m) \qquad (2.1$$

where w_{ij} can be viewed as the proportion contributed by $(L_i, R_i]$ in the estimation of p_j and hence $\sum_{j=1}^{m} w_{ij} = 1$.

We discuss the NPMLE for interval censored data and see how it relates to the self-consistency solution. The likelihood based on the original data can be written as

$$L = \prod_{i=1}^{n} F(R_i) - F(L_i) \qquad (2.1$$

which can be re-expressed as

$$L^* = \prod_{i=1}^{n}\left(\sum_{j=1}^{m}\delta_{ij}p_j\right)$$

where $0 \le p_j \le 1$ and $\sum_{j=1}^{m} p_j = 1$. Notice that the summation in L^* creates numerical difficulty in the maximization. The idea of EM is employed to solve the problem. If the variable $I_{ij} = I(T_i \in (l_j, r_j])$ can be known, the likelihood based on this "complete" information is given by

$$\tilde{L}^* = \prod_{i=1}^{n}\left(\prod_{j=1}^{m} p_j^{I_{ij}}\right)$$

and the corresponding log-likelihood is

$$\log \tilde{L}^* = \sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}\log p_j .$$

Maximization of $\log \tilde{L}^*$ subject to $\sum_{j=1}^{m} p_j = 1$ reduces to the multinomial problem. Applying the theorem of Lagrange multiplier, maximization is attained at

$$\hat{p}_j = \sum_{i=1}^{n} I_{ij} / n .$$

However I_{ij} is not known and the "E-step" imputes its value by

$$w_{ij}(p_1, ..., p_m) = \frac{\delta_{ij}p_j}{\delta_{i1}p_1 + ... + \delta_{im}p_m} .$$

Consequently the "M-step" maximizes

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}(p_1, ..., p_M)\log p_j . \qquad (2.1$$

Maximization is attained at $\hat{p}_j = \sum_{i=1}^{n} w_{ij}(\hat{p}_1,...,\hat{p}_m)/n$ which is equivalent to the self-consistency equation. The steps of iterations are summarized below.

Step 1: set $\hat{p}_j^{(0)} = 1/m \quad (j = 1,...,m)$;

Step 2: set $\hat{p}_j^{(k)} = \sum_{i=1}^{n} w_{ij}(\hat{p}_1^{(k-1)},...,\hat{p}_m^{(k-1)})/n$;

Step 3: repeat $k = 2,3,...$ (Step 2) for until convergence

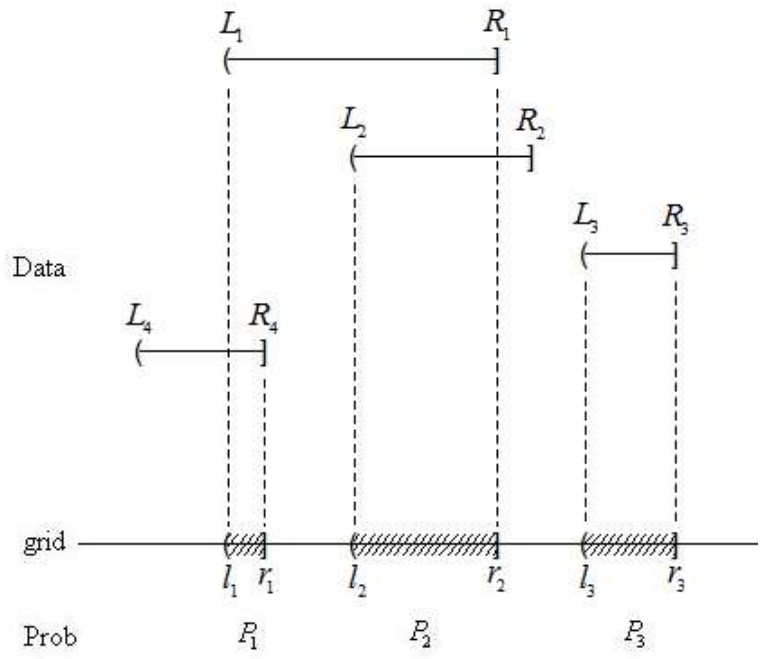


Figure 2.4 Construction of the mass interval censored data and the idea of self consistency. Weight for $(L_1, R_1]$ in estimating $\hat{P}_1, \hat{P}_2, \hat{P}_3$ are $1 \cdot P_1/(1 \cdot P_1 + 1 \cdot P_2 + 0 \cdot P_3)$ $1 \cdot P_2/(1 \cdot P_1 + 1 \cdot P_2 + 0 \cdot P_3)$ and 0 ,respectively.

A different algorithm which directly estimates the survival function was suggested by Turnbull (1976). His idea is based on the product-limit expression of

S(t). The hazard probability in the interval $(l_j, r_j]$ can be estimated by d_j / Y_j

where $Y_j = \sum_{k=j}^{k=m} d_k$ and $d_j = \sum_{i=1}^{n} \frac{\delta_{ij} p_j}{\sum_k \delta_{ik} p_k}$. The algorithm is given below.

Step 1: set $d_j^{(0)} = \sum_{i=1}^{n} \frac{\delta_{ij}}{\sum_k \delta_{ik}}$, $Y_j^{(0)} = \sum_{k=j}^{k=m} d_k^{(0)}$ $(j = 1, ..., m)$

Step 2: set $\hat{S}_j^{(l)} = \prod_{i \le j} \{1 - \frac{d_j^{(l)}}{Y_j^{(l)}}\}$;

Step 3: repeat $l = 2, 3, ...$ (Step 2) for until convergence.

Chapter 3. Inference based on Proportional Hazards Model

The proportional hazards model specifies the effect of covariate on the hazard of T such that

$$\lambda_Z(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} \Pr(T \in [t, t+\Delta) \mid T \geq t, Z)$$

$$= \lambda_0(t) \exp(Z^T \beta), \qquad (3.1)$$

where $\lambda_0(t)$ is an arbitrary baseline hazard rate and β is the vector of parameter . The survival function can be written as

$$S_Z(t) = \exp(-\int_0^t \lambda_Z(u)du) = \exp(-\int_0^t \lambda_0(u)\exp(Z^T\beta)du) = S_0(t)^{\exp(Z^T\beta)}$$

where $S_Z(t) = \Pr(T > t \mid Z)$. The main objective is to estimate β and usually estimation of $\lambda_0(t)$ is of less interest except for the purpose of prediction.

3.1 Inference under right censoring

Right censored data can be denoted as (X_i, δ_i, Z_i) $(i = 1, ..., n)$ where $X_i = T_i \wedge C_i$, $\delta_i = I(T_i \leq C_i)$, and Z_i is the covariate. Due to the expression of the likelihood,

$$L = \prod_{i=1}^{n} f_Z(x_i)^{\delta_i} S_Z(x_i)^{1-\delta_i} = \prod_{i=1}^{n} \lambda_Z(x_i)^{\delta_i} S_Z(x_i), \qquad (2.2)$$

the likelihood function under the PH model can be rewritten as

$$L(\beta, \lambda_0) = \prod_{i=1}^{n} \lambda_0(x_i)\exp(Z^T\beta)^{\delta_i} \exp(-\int_0^{x_i} \lambda_0(u)\exp(Z^T\beta)du) .$$

Note that under the semi-parametric setting, the parametric structure of $\lambda_0(t)$ is unspecified.

The partial likelihood approach based on right censored data allows one to estimate β without dealing with $\lambda_0(t)$. Let $t_{(1)}, ..., t_{(M)}$ be the ordered observed failure times. Let A_i be the set containing the information about who fails at $t_{(i)}$ and B_i be the set containing the censoring information in $[t_{(i)}, t_{(i-1)})$ and one failure

occurs at time $t_{(i)}$. The likelihood function can be expressed in terms of the sets $\{(A_j, B_j)(j = 1,...,M)\}$ such that

$$\begin{aligned}\tilde{L} &= \Pr(B_1 \cap A_1 \cap B_2 \cap A_2 \cap ... \cap B_M \cap A_M) \\ &= \prod_{j=1}^{M} \Pr(A_j \mid B^{(j)} \cap A^{(j-1)}) \Pr(B_j \mid B^{(j-1)} \cap A^{(j)}) \\ &= \prod_{j=1}^{M} \Pr(A_j \mid B^{(j)} \cap A^{(j-1)}) \prod_{j=1}^{M} \Pr(B_j \mid B^{(j-1)} \cap A^{(j)})\end{aligned}$$

where $A^{(j)} = (A_1 \cap ... \cap A_j)$ and $B^{(j)} = (B_1 \cap ... \cap B_j)$. Notice that

$$\begin{aligned}\Pr(A_j \mid B^{(j)}, A^{(j-1)}) &= \frac{\lambda_{z_{(j)}}(t_{(j)})dt_{(j)}}{\sum_{k \in R(t_{(j)})} \lambda_{z_k}(t_{(j)})dt_{(j)}} \\ &= \frac{\lambda_0(t_{(j)})\exp(Z_{(j)}^T \beta)dt_{(j)}}{\sum_{k \in R(t_{(j)})} \lambda_0(t_{(j)})\exp(Z_k^T \beta)dt_{(j)}} \\ &= \frac{\exp(z_{(j)}^T \beta)}{\sum_{k \in R(t_{(j)})} \exp(z_k^T \beta)}\end{aligned}$$

in which $\lambda_0(t)$ disappears in the last equation. It can be shown that ignoring the second component of $\tilde{L}$ still leads to an unbiased estimating equation. Specifically the partial likelihood

$$L_c(\beta) = \prod_{i=1}^{M} \Pr(A_j \mid B^{(j)} \cap A^{(j-1)}) = \prod_{j=1}^{M} \frac{\exp(Z_{(j)}^T \beta)}{\sum_{k \in R(t_{(j)})} \exp(Z_k^T \beta)} \qquad (3.3$$

which yields the score function:

$$U(\beta) = \frac{\partial}{\partial \beta} \log L_c = \sum_{j=1}^{M} (Z_{(j)} - \frac{\sum_{k \in R(t_{(j)})} \exp(Z_k^T \beta) Z_k}{\sum_{k \in R(t_{(j)})} \exp(Z_k^T \beta)}).$$

The estimator of β is the one which solves $U(\beta) = 0$. Breslow (1975) suggested to estimate $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ by

$$\Lambda_0(t) = \int_0^t \frac{\sum_{i=1}^{n} I(X_i = u, \delta_i = 1)}{\sum_{i=1}^{n} I(X_i \geq u)\exp(Z_i^T \hat{\beta})}. \qquad (3.4$$

3.2 Inference under interval censoring

Finkelstein (1986) was the first author to address the inference of proportional hazards models for interval censored data. Observed data can be written as (L_i, R_i, Z_i) $(i = 1, ..., n)$. It is assumed that T_i and (L_i, R_i) are independent given Z_i. The likelihood function can be written as

$$L = \prod_{i=1}^{n} [S_{Z_i}(L_i) - S_{Z_i}(R_i)] = \prod_{i=1}^{n} \left[\{S_0(L_i)\}^{Z_i^T \beta} - \{S_0(R_i)\}^{Z_i^T \beta} \right] \qquad (3.5$$

where $S_Z(t) = \Pr(T > t \mid Z)$. Unlike the case of right censoring, there is no way to remove $S_0(.)$ from (3.5). This means that we need to estimate β and $S_0(.)$ jointly. The likelihood function in (3.5) is represented based on original observations. For the purpose of maximization when the nuisance function $S_0(.)$ is involved, this function will be expressed in terms of grid points. Figure 3.1 shows the construction of grid points based on an example containing four observed intervals $(L_i, R_i]$ $(i = 1, 2, 3, 4)$. In this example, $t_{(1)} < t_{(2)} < \cdots < t_{(8)}$ are formed. In general, we let $0 = t_{(0)} < t_{(1)} < \cdots < t_{(M+1)} = \infty$ be ordered grid points. The i th observation to the likelihood (3.5) can be re-expressed as

$$\sum_{j=1}^{M+1} \delta_{ij} [S_{Z_i}(t_{(j-1)}) - S_{Z_i}(t_{(j)})] = \sum_{j=1}^{M+1} \delta_{ij} [S_0(t_{(j-1)})^{\exp(Z_i^T \beta)} - S_0(t_{(j)})^{\exp(Z_i^T \beta)}]$$

where $\delta_{ij} = 1$ if $(t_{(j-1)}, t_{(j)}] \subset (L_i, R_i]$ and 0 otherwise. Accordingly the likelihood in (3.5) can be written as

$$L = \prod_{i=1}^{n} \sum_{j=1}^{M+1} \delta_{ij} [S_0(t_{(j-1)})^{\exp(Z_i^T \beta)} - S_0(t_{(j)})^{\exp(Z_i^T \beta)}].$$

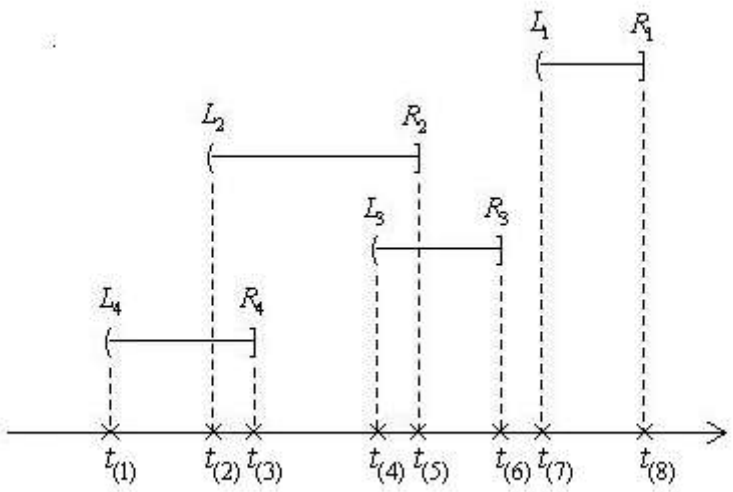


Figure 3.1 Construction of the mass interval censored data

Note that $S_0(.)$ is a function but the maximization of L will only be evaluated on its value at the grid points, namely $S_0(t_{(j)})$ for $j = 1,...,M$. Since the parameters $S_0(t_{(j)})$ for $j = 1,...,M$ have natural constraint i.e., $0 \le S_0(t_{(1)}) < \cdots < S_0(t_{(M)}) \le 1$, it make the mle hard to obtain. Finkelstein (1986) suggested that one can re-parametrize the $S_0(t_{(j)})$ by $\alpha_k = \log[-\log S_0(t_{(k)})]$. It can remove the range restriction of the parameters $S_0(t_{(j)})$, but it doesn't remove the order restriction.

Sun (2006) proposed a corrective method that can remove the nature constraint. The following is the detail of the method. Based on the product-limit expression, one can write

$$S_0(t_{(j)}) = \prod_{k=0}^{j} (1 - \Lambda_0(\Delta t_{(k)})) .$$

Thus L becomes a function of β and $\Lambda_0(\Delta t_{(1)}),...,\Lambda_0(\Delta t_{(M)})$. Note that $\Lambda_0(\Delta t_{(k)}) \in (0,1)$ is a conditional probability. To remove the natural constraint, we can re-parameterize $\Lambda_0(\Delta t_{(k)})$ by setting $\Lambda_0(\Delta t_{(k)}) = 1 - e^{-\exp(\alpha_k)}$.

Figure 3.2 shows that by the above transformation, $-\infty < \alpha_k < \infty$. Thus

$$S_0(t_{(j)}) = \prod_{k=0}^{j} e^{-\exp(\alpha_k)} = \mathrm{e}^{-\sum_{k=0}^{j}\exp(\alpha_k)}.$$

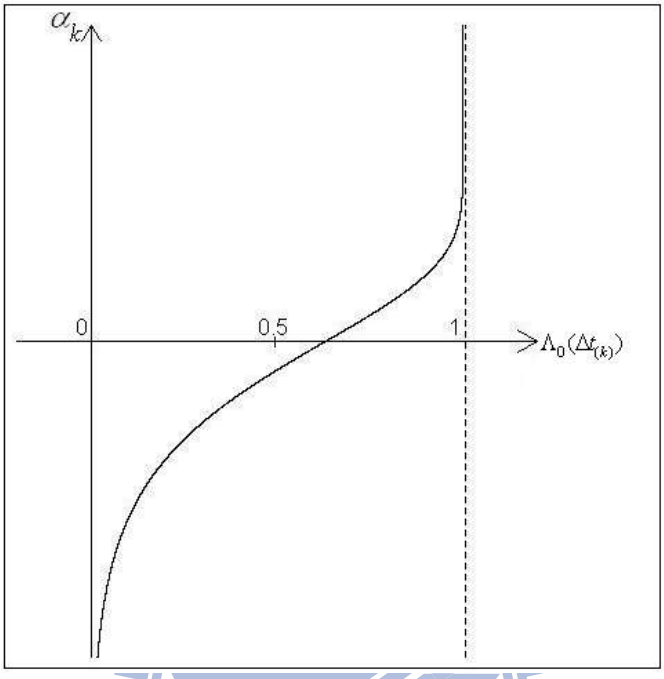


Figure 3.2 The relationship between $\Lambda_0(\Delta t_{(k)})$ and α_k

Finally, let

$$\gamma_j = -\sum_{k=0}^{j}\exp(\alpha_k),$$

for $j = 1,...,M$ and $\alpha_0 = -\infty$ and $\alpha_{M+1} = \infty$. We have obtained

$$L_k = \prod_{i=1}^{n}\sum_{j=1}^{M+1}\delta_{ij}\{\mathrm{exp}[\mathrm{exp}(Z_i^T\beta)\times\gamma_{j-1}] - \quad]\quad \mathrm{exp}(Z_i^T\beta)\mathrm{exp}(\qquad (3.6$$

Thus the log of likelihood becomes

$$l = \sum_{i=1}^{n}\log\sum_{j=1}^{M+1}\delta_{ij}\{\exp[\exp(Z_i^T\beta)\times\gamma_{j-1}] - \exp[\exp(Z_i^T\beta)\times\gamma_j]\}.$$

Note that the unknown parameters are $(\beta, \alpha_1, ..., \alpha_M)$, all of which have no additional constraints. The estimation process is summarized below.

Denote the score statistics as $U = (\frac{\partial l}{\partial \alpha^T}, \frac{\partial l}{\partial \beta^T})^T$. The score functions of α and β are given by

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} Z_i g_i^{-1} \sum_{j=1}^{M+1} \delta_{ij} (f_{i,j-1} - f_{i,j})$$

and

$$\frac{\partial l}{\partial \alpha_j} = \sum_{i=1}^{n} g_i^{-1} b_{i,j} c_{i,j} \quad (j = 1, ..., M)$$

where $g_i = \sum_{j=1}^{M+1} \delta_{ij} [S_{Z_i}(t_{(j-1)}) - S_{Z_i}(t_{(j)})]$, $f_{i,j} = S_{Z_i}(t_{(j)}) \log S_{Z_i}(t_{(j)})$, $f_{i,0} = f_{i,M+1} = 0$, $b_{i,j} = \exp(\alpha_j + Z_i^T \beta)$, $c_{i,j} = \sum_{k=j}^{M+1} (\delta_{ik} - \delta_{ik+1}) S_{Z_i}(t_{(k)})$ and $\delta_{i\,m+2} = 0$. The objective is to solve $U = 0$ so that the likelihood can be maximized. The Newton-Raphson method can be applied to obtain the solution. Let $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ as the observed Fisher information matrix. The elements of A_{11} include

$$\frac{\partial^2 l}{\partial \alpha_j \partial \alpha_k} = -\sum_{i=1}^{n} \left(\frac{b_{i,j} b_{i,k} c_{i,j} c_{i,k}}{g_i^2} + \frac{b_{i,j} b_{i,k} c_{i,j}}{g_i} \right) \quad \text{for} \quad j < k .$$

The elements of $A_{12} = A_{21}^T$ include

$$\frac{\partial^2 l}{\partial \alpha_j \partial \beta} = \sum_{i=1}^{n} Z_i b_{i,j} \{ g_i^{-1} [c_{i,j} + \sum_{l=j}^{M+1} (\delta_{il} - \delta_{il+1}) f_{i,l}] - \frac{c_{i,j}}{g_i^2} \sum_{l=1}^{M+1} \delta_{il} (f_{i,l-1} - f_{i,l}) \} .$$

The elements of A_{22} include

$$\frac{\partial^2 l}{\partial \beta \partial \beta^T} = \sum_{i=1}^{n} Z_i Z_i^T \left\{ \left[g_i^{-1} \sum_{j=1}^{M+1} \delta_{ij} (f_{i,j-1} - f_{i,j}) \right]^2 - g_i^{-1} \sum_{j=1}^{M+1} \delta_{ij} (h_{i,j-1} - h_{i,j}) \right\}$$

where

$$h_{i,j} = f_{i,j} \log S_{Z_i}(t_{(j)}) + f_{i,j} .$$

Denote X_n as the estimate of (α, β) in the n-iteration, the Newton-Raphson algorithm can be summarized as follow:

$$X_{K+1} = X_K + A_K^{-1} U_K \quad (K = 0, 1, ...),$$

where A_K and U_K are A and U evaluated at the Kth estimate of (α, β). The procedure is stopped when $X_{K+1} - X_K$ is close to zero.

Chapter 4. Inference based on Accelerated Failure Time Model

The accelerated failure time (AFT) is a popular alternative to the PH model. An AFT model can be written as

$$Y = Z^T\beta + \varepsilon \qquad (4.1$$

where $Y = \log T$ and ε is the error variable with the density function $f_\varepsilon(.)$. Existing research focuses on semi-parametric estimation of β without specifying the form of $f_\varepsilon(.)$. We will briefly review inference methods developed for complete data and right censored data and then focus more on interval censored data.

4.1 Inference based on complete data

Consider the transformed variable under the error scale, $\varepsilon_i(\beta) = Y_i - Z_i^T\beta$ for $i = 1,...,n$. Note that when β_0 is the true value, $\{\varepsilon_i(\beta_0)\ (i = 1,...,n)\}$ form an *iid* sample which becomes a useful property to construct nonparametric inference methods. Let $\varepsilon_{(1)}(\beta) < \varepsilon_{(2)}(\beta) < \cdots < \varepsilon_{(n)}(\beta)$ be the order statistics with the corresponding covariates denoted as $Z_{(1)}, Z_{(2)}, ..., Z_{(n)}$. Define c_i as the score of $Z_{(i)}$ satisfying $\sum_{i=1}^{n} c_i = 0$. Consider the linear rank statistic of the form:

$$v(\beta) = \sum_{i=1}^{n} Z_{(i)} c_i \,. \qquad (4.2$$

Note that when $\beta = \beta_0$, $\Pr(Z_{(i)} = Z_k) = 1/n$. As a result,

$$E(v(\beta_0)) = \sum_{i=1}^{n} c_i E(Z_{(i)}) = \sum_{i=1}^{n} c_i \bar{Z} = 0 \,.$$

This implies that one can estimate β by solving the equation $v(\beta) = 0$. The error distribution and the form of c_i both affect the efficiency of the resulting estimator. The Wilcoxon statistic corresponds to $c_i = 2i(n+1)^{-1} - 1$ and the log-rank statistic corresponds to $\sum_{j=1}^{i} n_j^{-1} - 1$ where n_j is the number at risk at $\varepsilon_{(j)}(\beta)$.

4.2 Inference based on right censored data

Right censored data are denoted as $\{(X_i, \delta_i, Z_i)\ (i = 1,2,...,n)\}$. Let $\tilde{\varepsilon}_i(\beta) = \tilde{Y}_i - Z_i^T \beta$, where $\tilde{Y}_i = \log(X_i)$. Censored data under the error scale can be written as $\{(\tilde{\varepsilon}_i(\beta), \delta_i)\ (i = 1,...,n)\}$. Note that $\{(\tilde{\varepsilon}_i(\beta_0), \delta_i)\ (i = 1,...,n)\}$ form a random sample of $\{\tilde{\varepsilon}(\beta_0), \delta\}$ which is a censored version of $\varepsilon(\beta_0)$. We discuss two methods of modifying the linear rank statistics introduced earlier.

4.2.1 Linear rank statistics

Let $\tilde{\varepsilon}_{(1)}(\beta) < \tilde{\varepsilon}_{(2)}(\beta) < \cdots < \tilde{\varepsilon}_{(k)}(\beta)$ be the distinct uncensored ordered error variables and let $\tilde{\varepsilon}_{i1}(\beta),...,\tilde{\varepsilon}_{im_i}(\beta)$ be censored error variables in $\left[\varepsilon_{(i)}(\hat{\beta}), \varepsilon_{(i+1)}(\hat{\beta})\right)$. Following the book of Kalbfleisch and Prentice (2002), the modified linear rank statistics can be written as

$$v(\beta) = \sum_{i=1}^{k} (c_i Z_{(i)} + \sum_{j=1}^{m_i} C_i Z_{ij}) \qquad (4.3$$

where $Z_{(i)}$ and Z_{ij} represent the corresponding covariates of $\tilde{\varepsilon}_{(i)}(\beta)$ and $\tilde{\varepsilon}_{ij}(\beta)$ respectively. The requirement is given below:

$$\sum_{i=1}^{k} (c_i + m_i C_i) = 0 .$$

The efficiency of the resulting estimator depends on the underlying error distribution, the censoring distribution and the forms of c_i and C_i . For the Wilcoxon statistics, Prentice (1978) suggested that

$$c_i = 1 - 2\prod_{j=1}^{i} \frac{n_j}{1+n_j} \quad \text{and} \quad C_i = 1 - \prod_{j=1}^{i} \frac{n_j}{1+n_j},$$

where n_j is the number at risk at $\tilde{\varepsilon}_{(j)}(\beta)$.

4.2.2 Log-rank statistics

As mentioned earlier the log-rank statistics can be expressed as a special linear rank statistics but it has its own representation as follows:

$$U^{LR}(\beta)=\sum_i \delta_i\left[Z_i-\frac{\sum_{j=1}^{n} I(\tilde{\varepsilon}_j(\beta)\geq \tilde{\varepsilon}_i(\beta))Z_j}{\sum_{j=1}^{n} I(\tilde{\varepsilon}_j(\beta)\geq \tilde{\varepsilon}_i(\beta))}\right]. \qquad (4.4$$

4.2.3 M-estimator by Ritov (1990)

The idea was motivated by the parametric likelihood analysis. The likelihood function and log-likelihood based on $(\tilde{\varepsilon}_i(\beta),\delta_i)$ $(i=1,...,n)$ can be written as

$$L(\beta, f_\varepsilon)=\prod_{i=1}^{n} f_\varepsilon(\tilde{\varepsilon}_i(\beta))^{\delta_i} S_\varepsilon(\tilde{\varepsilon}_i(\beta))^{1-\delta_i} \qquad (4.5$$

and

$$l(\beta, f_\varepsilon)=\sum_{i=1}^{n}\delta_i \log f_\varepsilon(\tilde{\varepsilon}_i(\beta))+(1-\delta_i)\log S_\varepsilon(\tilde{\varepsilon}_i(\beta))$$

respectively. Taking derivative respect to β, we get

$$\frac{\partial l(\beta, f_\varepsilon)}{\partial \beta}=\sum_{i=1}^{n}[-\delta_i\frac{f_\varepsilon'(\tilde{\varepsilon}_i(\beta))}{f_\varepsilon(\tilde{\varepsilon}_i(\beta))}Z_i(\beta)+(1-\delta_i)\frac{f_\varepsilon(\tilde{\varepsilon}_i(\beta))}{S_\varepsilon(\tilde{\varepsilon}_i(\beta)}Z_i(\beta)].$$

The difficulty comes from the fact that f_ε, f_ε' and S_ε are all unknown functions. The idea of M-estimator is to replace $\frac{-f_\varepsilon'(\cdot)}{f_\varepsilon(\cdot)}$ by $g(\cdot)$ which is a known function. Accordingly $\frac{\partial l(\beta, f_\varepsilon)}{\partial \beta}$ can be replaced by

$$U^M(\beta, S_\varepsilon)=\sum_{i=1}^{n} Z_i[\delta_i g(\tilde{\varepsilon}_i(\beta))+(1-\delta_i)\frac{-\int_{\tilde{\varepsilon}_i(\beta)}^{\infty} g(u)dS_\varepsilon(u)}{S_\varepsilon(\tilde{\varepsilon}_i(\beta))}] \qquad (4.6$$

where $S_\varepsilon(\cdot)$ can be estimated by the following product-limit estimator

$$\hat{S}_\varepsilon(t;\beta)=\prod_{u\leq t}\left\{1-\frac{\sum_{j=1}^{n} I(\tilde{\varepsilon}_j(\beta)=u,\delta_j=1)}{\sum_{j=1}^{n} I(\tilde{\varepsilon}_j(\beta)\geq u)}\right\}.$$

The form of $g(.)$ and the underlying error distribution affect the efficiency of

the resulting estimator. Two common choices of $g(\cdot)$ are $g(u)=u$ or

$$g(u)=\begin{cases} -k & if\ u<-k \\ u & if\ |u|\le k \\ k & if\ u>k. \end{cases}$$

Note that if the error distribution is specified, we have

$$f_\varepsilon=-\int_t^\infty f_\varepsilon'(x)dx=-\int_t^\infty \frac{f_\varepsilon'(x)}{f_\varepsilon(x)} f_\varepsilon(x)dx=-\int_t^\infty g(x)dF_\varepsilon(x)$$

which suggests the best form of $g(.)$ is $\frac{f_\varepsilon'(.)}{f_\varepsilon(.)}$ since it yields the maximum likelihood estimator.

4.3 Inference under interval censoring

Consider interval censored data (L_i,R_i,Z_i) $(i=1,...,n)$ which can be transformed under the error scale such that $\varepsilon_i^L(\beta)=\log L_i-Z_i^T\beta$ and $\varepsilon_i^R(\beta)=\log R_i-Z_i^T\beta$. Now we discuss how to extend the ideas of previous methods to interval censored data.

4.3.1 Modify M-estimator by Rabinowitz et al. (1995)

Consider the log-likelihood function

$$L=\sum_{i=1}^n \ln[F_\varepsilon\{\varepsilon_i^R(\beta)\}-F_\varepsilon\{\varepsilon_i^L(\beta)\}].$$

The MLE of β can be obtained by solving the score function

$$S(\beta)=\sum_{i=1}^n \frac{f_\varepsilon\{\varepsilon_i^R(\beta)\}-f_\varepsilon\{\varepsilon_i^L(\beta)\}}{F_\varepsilon\{\varepsilon_i^R(\beta)\}-F_\varepsilon\{\varepsilon_i^L(\beta)\}} Z_i \qquad (4.7$$

and it follows that $E(S(\beta_0))=0$. Since the forms of f_ε and F_ε are un-specified, one can replace the middle part of $S(\beta)$ by

$$c_i(\beta,F_\varepsilon)=\frac{g[F_\varepsilon\{\varepsilon_i^R(\beta)\}]-g[F_\varepsilon\{\varepsilon_i^L(\beta)\}]}{F_\varepsilon\{\varepsilon_i^R(\beta)\}-F_\varepsilon\{\varepsilon_i^L(\beta)\}}$$

where the function g with domain [0,1] satisfies $g(0)=g(1)=0$. Thus one can estimate β by solving

$$\tilde{S}(\beta, F_\varepsilon) = \sum_{i=1}^{n} c_i(\beta, F_\varepsilon) Z_i = 0 .$$

In Appendix, we show why the restrictions on g makes $E[\tilde{S}(\beta_0, F_\varepsilon)] = 0$. Again the form of $g(.)$ and the underlying error distribution affect the efficiency of the resulting estimator. The best choice which leads to the maximum likelihood estimator is $g = f_\varepsilon \circ F_\varepsilon^{-1}$.

Recall that in Section 4.2.3, the nuisance function S_ε has an explicit product-limit estimator and hence $U^M(\beta, \hat{S}_\varepsilon(.|\beta)) = 0$ is solved. However for interval censored data, there is no explicit formula for estimating S_ε or F_ε. One way of estimating F_ε is by maxing the likelihood based on $\{(\varepsilon_i^L(\beta), \varepsilon_i^R(\beta))(i = 1,...,n)\}$. The grid intervals in which F_ε receives masses can be constructed similar to the setup in Section 2.4. Denote $t_k(\beta) = (t_{kL}(\beta), t_{kU}(\beta)]$ for $k = 1,...,m$ as the mass intervals. The next step is to express the likelihood based on $\{(\varepsilon_i^L(\beta), \varepsilon_i^R(\beta))(i = 1,...,n)\}$ in terms of $\{(t_{kL}(\beta), t_{kU}(\beta)], k = 1,...,m\}$. Define $\tilde{t}_{i,L}(\beta) = \max\{t_{j,L}(\beta) : t_{j,L}(\beta) \le \varepsilon_i^L(\beta)\}$ and $\tilde{t}_{i,U}(\beta) = \min\{t_{jR}(\beta) : t_{jR}(\beta) \ge \varepsilon_i^R(\beta)\}$. That is $(\tilde{t}_{i,L}(\beta), \tilde{t}_{i,R}(\beta)]$ becomes a new representation of $(\varepsilon_i^L(\beta), \varepsilon_i^R(\beta)]$ which will be a union of consecutive $t_k(\beta)$. The corresponding log-likelihood function becomes

$$l = \sum_{i=1}^{n} \ln[F_\varepsilon\{\tilde{t}_{i,U}(\beta)\} - F_\varepsilon\{\tilde{t}_{i,L}(\beta)\}]$$

subject to

$$F_\varepsilon\{\tilde{t}_{i,U}(b)\} - F_\varepsilon\{\tilde{t}_{i,L}(b)\} \in [0,1] \ \ (i = 1,2,...,n)$$

and

$$F_\varepsilon\{t_j(\beta)\} \in [0,1] \ \ (j = 1,2,...,m) .$$

Given an estimate of β, the above maximization procedure can be performed. The resulting estimate of F_ε, denoted as $\hat{F}_\varepsilon$, will be plugged into the following estimating function to solve for the next estimate of β:

$$\tilde{S}(\beta, \hat{F}_\varepsilon) = \sum_{i=1}^{n} c_i(\beta, \hat{F}_\varepsilon)Z_i = \quad . \qquad (4.8$$

4.3.2 Method for a simplified AFT model with univariate covariate

Besides the methods we discussed earlier, Li and Pu (2003) developed an interesting way to estimate β. Consider a simplified AFT model:

$$Y_i = \beta_i Z_i + \varepsilon_i, \qquad (i = 1, \ldots n).$$

where Z_i is a one-dimensional covariate. The main idea of this paper is based on the assumption that ε_i and Z_i are uncorrelated. Kendall's provides a rank-invariant measure for assessing the association between two variables. Suppose that (ε_i, Z_i) and (ε_j, Z_j) are independent realizations from (ε, Z). The pair is *concordant* if $I\{(\varepsilon_i - \varepsilon_j)(Z_i - Z_j) > 0\}$ and *discordant* if $I\{(\varepsilon_i - \varepsilon_j)(Z_i - Z_j) < 0\}$. The population version of Kendall's tau is defined as

$$\tau = \Pr\{(\varepsilon_i - \varepsilon_j)(Z_i - Z_j) > 0\} - \Pr\{(\varepsilon_i - \varepsilon_j)(Z_i - Z_j) < 0\}.$$

The sample estimate of τ is

$$\overline{\tau} = \frac{\sum_{i<j} I\{(\varepsilon_i - \varepsilon_j)(Z_i - Z_j) > 0\} - I\{(\varepsilon_i - \varepsilon_j)(Z_i - Z_j) < 0\}}{\binom{n}{2}/2}.$$

If complete data are available, one can solve

$$\frac{1}{n(n-1)} \sum_{i<j} I\{(\varepsilon_i(\beta) - \varepsilon_j(\beta))(Z_i - Z_j) > 0\} - I\{(\varepsilon_i(\beta) - \varepsilon_j(\beta))(Z_i - Z_j) < 0\} = 0$$

to estimate β. However $\varepsilon_i(\beta) = Y_i - \beta_i Z_i$ is subject to interval censoring such that we only know that $\varepsilon_i(\beta) \in (\varepsilon_i^L(\beta), \varepsilon_i^R(\beta)]$. Some interval observations provide the

complete information for the order of $\varepsilon_i(\beta)$ and $\varepsilon_j(\beta)$. Notice that if

$\varepsilon_i^R(\beta) < \varepsilon_j^L(\beta)$, then $\varepsilon_i < \varepsilon_j$.

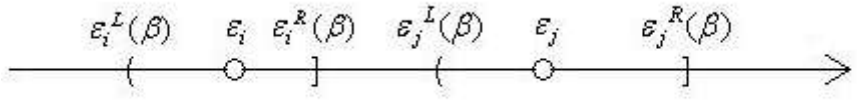


Figure 4.1 The relative position of ε_i and ε_j when $\varepsilon_i^R(\beta) < \varepsilon_j^L(\beta)$

Hence the total number of known concordant pairs becomes

$$\sum_{i<j}\{I(Z_i < Z_j)I(\varepsilon_i^R(\beta) < \varepsilon_j^L(\beta)) + I(Z_i > Z_j)I(\varepsilon_i^L(\beta) > \varepsilon_j^R(\beta))\}$$

and the total number of known discordant pairs is

$$\sum_{i<j}\{I(Z_i < Z_j)I(\varepsilon_i^L(\beta) > \varepsilon_j^R(\beta)) + I(Z_i > Z_j)I(\varepsilon_i^R(\beta) < \varepsilon_j^L(\beta))\}.$$

The modified Kendall τ coefficient can be write as

$$\frac{1}{n(n-1)}\sum_{i<j}[I(Z_i < Z_j) - I(Z_i > Z_j)][I(\varepsilon_i^R(\beta) \le \varepsilon_j^L(\beta)) - I(\varepsilon_i^L(\beta) > \varepsilon_j^R(\beta))].$$

The resulting estimation function is given by

$$K_n(\beta) = \sum_{i<j}[I(Z_i < Z_j) - I(Z_i > Z_j)][I(\varepsilon_i^R(\beta) < \varepsilon_j^L(\beta)) - I(\varepsilon_i^L(\beta) > \varepsilon_j^R(\beta))]$$

This method has two major drawbacks. One is the restrictive assumption that Z is univariate. The other is the lack of efficiency if the data contains very few orderable paired intervals.

Chapter 5. Inference based on Proportional Odds Model

5.1 Model and the likelihood

Besides the PH and AFT models, the proportional odds (PO) model is also a popular choice. The proportional odds model is defined as

$$\text{logit}\{F_Z(t \mid Z)\} = \log\{B(t)\} + \beta' Z \quad (5.1$$

where $\text{logit}(t)=\log\{t/1-t\}$, and $F_Z(t) = P(T \le t \mid Z)$. Note that B(t) is a non-decreasing baseline function with $B(0) = 0$. Accordingly the distribution and density functions become

$$F_Z(t) = \frac{B(t)e^{-\beta' Z}}{1 + B(t)e^{-\beta' Z}}$$

and

$$f_Z(t) = \frac{B^{(1)}(t)e^{-\beta' Z}}{\{1 + B(t)e^{-\beta' Z}\}^2}$$

respectively where $B^{(1)}(t)$ is the derivative of $B(t)$. The parameter of interest is β and $B(t)$. As before, we will examine likelihood-based inference methods.

Under right censoring, the observations consist of $\{(X_i, \delta_i, Z_i)\ (i = 1,2,...,n)\}$. The log-likelihood function is

$$l(\theta) = n^{-1}\sum_{i=1}^{n} \delta_i \log f_Z(x_i) + (1-\delta_i)\log F_Z(x_i) \quad (5.2$$

where $\theta = (\beta, B(t))$. Under interval censoring, the observed data consist of $\{(L_i, R_i, Z_i)\ (i = 1,2,...,n)\}$. The log-likelihood function becomes

$$l(\theta) = n^{-1}\sum_{i=1}^{n} \log\{F_{Z_i}(R_i) - F_{Z_i}(L_i)\} \quad (5.3$$

The presence of $B(t)$ makes it difficult to directly obtain the M.L.E. by maximizing the log-likelihood function. We will present two methods both of which suggested to

replace $B(t)$ by an approximated function which is easier to handle.

5.2 Smoothing method for approximating the baseline function

Shen (1998) proposed a sieve method to approximate $B(t)$ in the likelihood function. The method can be applied to not only right censored data but also interval censored data. Now we briefly describe the approach. The basic idea is that the baseline $B(t)$ can be approximated by a parametric function. Specifically define $I = (0 < t_{(1)} < \cdots < t_{(k)} < \infty)$ be the location of knots and $M = (m_0 < m_1 < \cdots < m_k)$ as the order of polynomial in these subintervals. Here $m_i < N_{\max}$, $k < K$ where $N_{\max}$ and K are pre-assigned numbers. For convenience, we define $h = (I, M)$. Let

$$s(t) = \sum_{i=0}^{k} \sum_{j=0}^{m_i} \eta_{ij} t^j I(t_{(i)} < t \le t_{(i+1)}) \qquad (5.4$$

as splines with variable orders and knots. An example of $s(t)$ is depicted in Figure 5.1. There are two knots which form three intervals. Each interval contains a polynomial of different orders. From this figure, we see that $s(t)$ can be used to approximate any smooth functions.

The approximated function of $B(t)$ is defined as

$$\hat{B}(t) = \int_o^t \exp(s(x))dx$$

In order to ensure the smoothness of $\hat{B}(t)$, the function $s(t)$ must satisfy the following constraint: The spline at $t_{(i)}$ have a derivative of order $\min(m_{i-1}, m_i)$. Thus, we transform the parameters $\theta = \{B(t), \beta\}$ into $\tilde{\theta} = \{\eta = (\eta_{ij}), \beta\}$ which can be estimated from the log-likelihood $l(\tilde{\theta})$ subject to the constraint.

The choice of h is based on its comparative Kullback-Leibler risk $R(h)$ which is defined as $-E_\theta(l(\hat{\theta}))$ where $\hat{\theta}$ is the estimator of θ . Since the parameter θ is

unknown, we cannot obtain it directly. For each fixed h, let $\hat{\theta}_i$ be the sieve maximum likelihood estimate of θ without the ith observation, and $\hat{P}_{-i}(.)$ is the corresponding estimated distribution. Shen (1998) suggests that we can use the statistics

$$-n^{-1}\sum_{i=1}^{n}\log \hat{P}_{-i}(\Omega_i)$$

to estimate $-E_\theta(l(\hat{\theta}))$ where Ω_i is the ith observation. This is the selector value of h. Then we choose optimal h that minimizes $R(h)$.

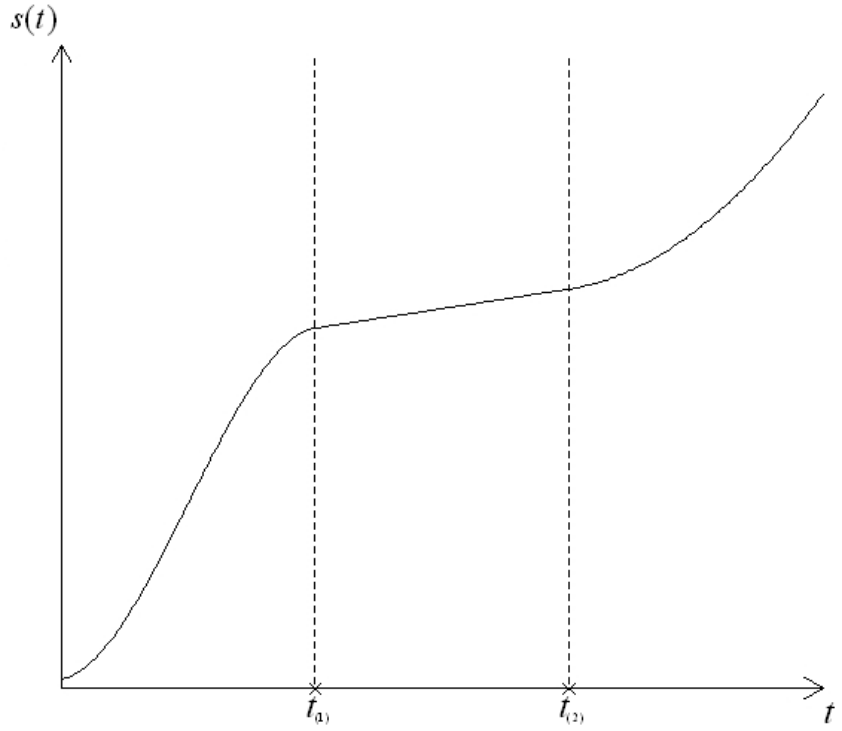


Figure 5.1. An example of $s(t)$. Here $t_{(1)}=4$ and $t_{(2)}=8$. The polynomials from left to right are $-x^3+6x^2+x-24$, $x+8$ and $x^2-15x+72$, respectively.

Hence we find the universal sieve maximum likelihood estimator by estimating $\tilde{\theta}$ and h recursively. The detail of the algorithm is as following.

Step 1: Initial spline

For any fixed order $m < N_{\max}$, estimate $\{\beta, \eta\}$ using the maximum likelihood

method with single polynomial. Then we choose the optimal order m that minimizes $R(h)$

Step 2: Adding knots

Consider a candidate knot point $t_{(i)}$ within an interval spanned by existed knots. For any fixed order $M=(m_0,m_1)$, estimate $\{\beta,\eta\}$ as step 1. Find the order M that minimize $R(h)$. This value is the selector of this candidate. Then the optimal $t_{(i)}$ is found using Fibonacci search to minimize the selector.

Step3: Comparison

Compare the original sieve maximum likelihood estimate based on the spline without $t_{(i)}$ with the new one including $t_{(i)}$. If the new maximum likelihood estimate has a smaller value in terms of the selector, then split the interval into two and proceed further as in Step 2. Otherwise, go to Step 4.

Step4: Repeat Steps 2-3 for all intervals spanned by existing knots until no new knot can be added

5.3 Sieve method by Huang and Rossini (1997)

The proportional odds model is expressed as

$$\log it F_Z(t \mid Z) = \log it F_o(t) + \beta^T Z \qquad (5.5$$

where $F_0(t)=F(t\mid 0)$ is the baseline distribution function. Let $\alpha_0(t)=\log it\, F_o(t)$, the distribution function can be written as

$$F_Z(t)=\frac{\exp(\alpha_0(t)+\beta^T Z)}{1+\exp(\alpha_0(t)+\beta^T Z)}$$

The difficulty of estimating β comes from the presence of $\alpha_0(t)$. Huang and Rossini (1997) proposed to is difficult to estimate this function by a function with nice analytic properties. The idea of this approximation is similar to the previous method.

If the real function $\alpha_0(t)$ is known, we might choose some knots $0 \le t_{(1)} < \cdots < t_{(k)} < \infty$ and let

$$\hat{\alpha}_0(t) = \sum_{j=1}^{k} \left[\frac{b_j - b_{j-1}}{t_{(j)} - t_{(j-1)}} t - \frac{b_j t_{(j-1)} - b_{j-1} t_{(j)}}{t_{(j)} - t_{(j-1)}} \right] I[t_{(j-1)} \le t \le t_{(j)}] \qquad (5.6)$$

where $b_j = \alpha_0(t_{(j)})$ and $b_1 \le \cdots \le b_k$. Here we choose k and $t_{(i)}$ which satisfy

1. k be an integer that grows at rate $O(n^a)$ $0 < a < 1$
2. $\max_{1 \le j \le k} (t_{(j)} - t_{(j-1)}) \le Cn^{-a}$ for some constant C

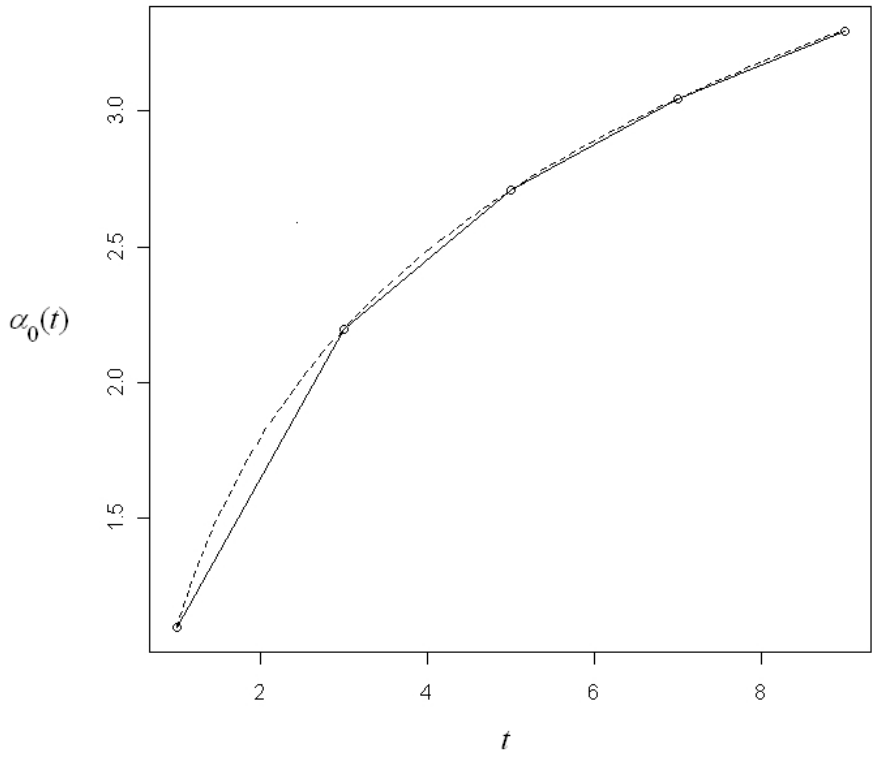


Figure 5.2. The curve of $\alpha_0(t)$ (dashed line) and its approximate function(real line). Here $\alpha_0(t) = \log(3t)$ and we take $t_{(.)} = 1,3,5,7,9$

There is some difficulty to implement the idea. Since the true function is unknown, $b_j = \alpha_0(t_{(j)})$ is also unknown. Treating b_j as unknown, the restriction that $b_1 \le \cdots \le b_k$ has to be considered in the maximization. The estimator of Shen (1998) is easier to implement since the unknown parameters have no specific restrictions.

Chapter 6 Conclusion

Most textbooks on survival analysis focus on right censored data. However empirical medical data are often interval censored. In this thesis, we review important inference methods which can be applied to interval censored data. We emphasize how the fundamental ideas of inference are extended to this complicated data structure. From the discussions, we see that many elegant techniques adopted for right censored data no longer applied. Instead, numerical algorithms become very important in analysis of interval censored data. Because the main purpose of the thesis is to provide a general review of many different methods, we do not investigate thoroughly on specific methods or algorithms. This can be interesting topics for future study.

References

[1] Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *Internat. Statist. Rev*., **43**, 45-58

[2] Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**, 845–854.

[3] Huang, J. and Rossini, A. J. (1997). Sieve estimation for the proportional odds failure-time regression model with interval censoring. *J. Amer. Statist. Ass*., **92**, 960-967.

[4] Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.*, **53**, 457-81.

[5] Kalbfleisch, J. D. and Prentice, R. L. (2002). Statistical analysis of failure time data, 2nd ed. New York. Wiley

[6] Li, L. and Pu, Z. (2003). Rank estimation of log-linear regression with interval censored data. *Lifetime Data Analysis*, **9**, 57–70.

[7] Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, **65**, 167-179.

[8] Rabinowitz , D., Tsiatis, A. and Aragon, J. (1995). Regression with interval censored data. *Biometrika*, **82**, 501-513.

[9] Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist*, **18,** 303-328

[10] Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika*, **85**, 165-177

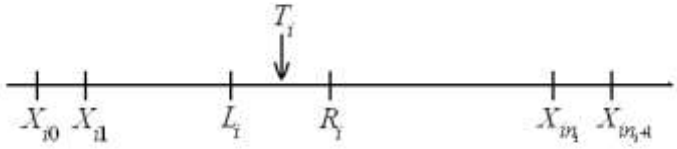
[11] Sun, J. (2006). The statistical analysis of interval-censored failure time data. USA. Springer

[12] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Ass*., **69**, 169-173.

[13] Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped censored and truncated data. *J. Roy. Statist. Soc. Ser*. *B*, **38**, 290– 295.

Appendix

Proof of $\mathbf{E[\tilde{S}(\beta_0, F_\varepsilon)] = 0}$. Let $\tilde{X}_i = (X_{i0} < X_{i1} < \cdots < X_{i\,n_i+1})$ be the ith patient's ordered sequence of examination times, where n_i denote the number of examination. For convenience. define $X_{i0} = -\infty$, and $X_{i\,n_i+1} = \infty$. Define L_i be the last of the ith subject's examination times preceding T_i, and let R_i be the first examination time following T_i.



For a p-dimensional vector b, define bracketing examination times on the time scale of the residual by

$$\varepsilon_i^{\,L}(b) = \log L_i - Z_i^t b\,, \quad \varepsilon_i^{\,R}(b) = \log R_i - Z_i^t b$$

We only have to prove

$$E(c_i(b)\mathrm{Z}_i \mid \tilde{X}_i, Z_i) = Z_i E\left(\frac{g[F\{\varepsilon_i^{\,R}(b)\}] - g[F\{\varepsilon_i^{\,L}(b)\}]}{F\{\varepsilon_i^{\,R}(b)\} - F\{\varepsilon_i^{\,L}(b)\}} \mid \tilde{X}_i, Z_i\right) = 0$$

where function g with domain [0,1] satisfies $g(0) = g(1) = 0$, then

$$E[\tilde{S}(\beta_0, F_\varepsilon)] = \sum E(c_i(b)\mathrm{Z}_i \mid \tilde{X}_i, Z_i) = 0$$

Consider

$$P(X_{i\,k} < T_i \le X_{i\,k+1} \mid \tilde{X}_i, Z_i)$$

$$= P(X_{i\,k} - Z_i^t b < T_i \le X_{i\,k+1} - Z_i^t b \mid \tilde{X}_i, Z_i)$$

$$= F(X_{i\,k+1} - Z_i^t b) - F(X_{i\,k+1} - Z_i^t b)$$

$$= F(\varepsilon_{i\,k+1}(b)) - F(\varepsilon_{i\,k}(b))$$

Thus,

$$Z_i E\left(\frac{g[F\{\varepsilon_i^{\,R}(b)\}] - g[F\{\varepsilon_i^{\,L}(b)\}]}{F\{\varepsilon_i^{\,R}(b)\} - F\{\varepsilon_i^{\,L}(b)\}} \mid \tilde{X}_i, Z_i\right)$$

$$= \sum_{k=0}^{n_i} \frac{g[F\{\varepsilon_{ik+1}(b)\}] - g[F\{\varepsilon_{ik}(b)\}]}{F\{\varepsilon_{ik+1}(b)\} - F\{\varepsilon_{ik}(b)\}} P(X_{i\,k} < T_i \le X_{i\,k+1} \mid \tilde{X}_i, Z_i)$$

$$= \sum_{k=0}^{n_i} g[F\{\varepsilon_{ik+1}(b)\}] - g[F\{\varepsilon_{ik}(b)\}]$$

$$= \sum_{k=0}^{n_i} g[F\{\varepsilon_{ik+1}(b)\}] - g[F\{\varepsilon_{ik}(b)\}]$$

$$= g[F\{\varepsilon_{i\,n_i+1}(b)\}] - g[F\{\varepsilon_{i0}(b)\}]$$

$$= g[F\{\infty\}] - g[F\{-\infty\}]$$

$$= g[1] \; - g[0] \quad = 0$$

The proof is complete.