國立臺灣大學理學院數學系
碩士論文
Department of Mathematics
College of Science
National Taiwan University
Master Thesis

申訴理賠時間和醫療成本的二元分配函數
Bivariate Distribution of Claiming Time and Medicare Reimbursement Based on Incomplete Data

黃姿蓉
Tzu-Jung Huang

指導教授: 江金倉 博士
Advisor: Chin-Tsang Chiang, Ph.D.

中華民國九十七年七月
July, 2008

口試委員審定書

Table of Contents

List of Tables

List of Figures

Acknowledgements

感謝江金倉老師的指導與啓發, 讓我從對做研究一無所知, 到能對生物統計相關的議題都有大致的了解。老師對做研究的熱誠與堅持, 讓人十分地感動。這篇論文的完成, 要感謝的人太多了, 以至於我未能一一細數。許多加油的隻字片語, 出現在不同的字裡行間; 更多令人會心溫暖的影像, 紛然雜陳。已忘記當初確切是如何的動力, 引領我走至今時今日, 就如同每一滴水, 都對於即將傾盆的暴雨有所貢獻。然而在每一場大雨後, 我們總是都看不清水滴的面目。唯一能說的是, 未來不論我在何處何地, 我都不會忘記身爲一個統計人的快樂, 驕傲與責任:
”The joy and job of statisticians are to listen to carefully to the data, not to tell the data how to behave.”

最後, 僅將此論文獻給, 總是無怨無悔支持著我, 最最親愛的爸爸與媽媽。

摘要

目前有關二元分配函數分析的研究, 都集中於處理雙存活時間的資料結構, 以及探討其相關的統計推論。有別於目前大部分的研究主題, 在這篇論文中, 我們關心存活時間(申訴理賠時間)和指標變數(申訴產生的醫療理賠) 的二元分配函數。根據實際觀測的右設限資料, 我們利用機率倒數加權法(Inverse Probability Weighting Method) 和替換法 (Imputation Method) 提出了更具延伸性的估計式。同時, 我們也運用機率倒數加權法, 針對有終點事件 (Terminal Events) 發生的情況, 提出相關的二元分配函數的估計式。進一步, 我們更建立上述估計式的大樣本性質, 利用估計式的高斯過程逼近, 配合著變異矩陣的估計式, 建構其相對應的信賴區間。我們執行了一系列的模擬檢證這些估計式以及信賴區間在有限樣本下之特性, 此外, 應用所提出的估計法在 SEER-Medicare 資料庫有關結腸癌的資料上。

關鍵字: 二元分配函數; 指標變數; 醫療成本; 設限; 終點事件; 機率倒數加權法; 替換法; 高斯過程

Abstract

Although numerous attempts have been made on bivariate failure times, however, there is little attention on the study of relation between failure time(claiming time) and mark variable(medicare reimbursement). Meanwhile, from the viewpoints of brokers and the management in insurance company, it is more attractive and engrossing to capture the dynamic pattern so that our research interest would focus on the joint distribution of claiming time and the corresponding medicare reimbursement. Based on survival censored data, we propose two estimation procedures: the inverse probability weighting (IPW) method and the imputation method. Furthermore, it is meaningful to accommodate terminal events occurring prior to the realization of failure time and the IPW method could lead to the resolution of this obstacle. Moreover, the limiting Gaussian processes of estimated distributions and the estimators of variance-covariance functions are developed and enable us to construct approximated regions. To investigate the finite sample properties of proposed estimators and the performance of inference procedures, a class of simulations would be conducted. An application to the colorectal cancer data retrieved from the Surveillance, Epidemiology, and End Results (SEER) Medicare database is also presented. In the end, we provide a brief discussion and further research topics of interest.

KEY WORDS: bivariate distribution, mark variable, medical cost, censoring, terminal events, inverse probability weighting (IPW), imputation, U-statistics, Gaussian processes

Chapter 1

Introduction

Bivariate failure time data are often encountered in many biomedical contexts, for example, times to blindness in both eyes of diabetic retinopathy patients or gap times to successive stages in the progression of acquired immunodeficiency syndrome. Estimation methods for dealing with parallel bivariate survival times can be found in Campbell (1981), Tsai, et al. (1986), Burke (1988), Dabrowska (1988), and Akritas, et al.(2003), among others. Statistical analyses for bivariate serial gap times include the works of Wang, et al. (1998) and Lin, et al. (1999). Distinguished from bivariate failure time data, an overwhelming emphasis is placed on health care cost controlling and fiscal accountability in medicine, and a brand new survival-type data structure emerge, called the cost data. The cost data consist of claiming time (time to the occurrence of medicare reimbursement) and the mark variable (medical cost or reimbursement), which is possibly correlated to claiming time and not observed until the corresponding claiming time fulfills. Rather than censoring, the marker would be missing due to the loss of follow-up or drop-out. Thus, it is inappropriate to use the existing statistical methods for bivariate failure time data.

Even though medical cost management has received immense attention, researchers merely highlighted the topic in two primary aspects over the past decades. One is mean cost evaluation within the period of interest based on complete or censored data. The other is cost-effectiveness (CE) ratio, the ratio of the mean difference in costs to the mean difference in effectiveness. CE ratio is a practical measurement for the trade off between budget constraints and patient benefits. In contrast, from the viewpoints of brokers and the management in insurance company, it is more crucial and appealing to explore the joint distribution of the claiming time X^o versus the medicare reimbursement Y^o. Few attempts have made at this sort of phenomena except for Huang and Louis (1998) and Hudgens, et al (2007). Based on the censored cost data $\{X_i,\ \delta_i,\ Y_i\}_{i=1}^n$ with $X_i = X_i^o \wedge C_i,\ Y_i = \delta_i Y_i^o,\ \delta_i = I(X_i^o \leq C_i)$ being the censoring status, and C being the censoring time, Huang and Louis obtained the estimator for joint distribution $F(t,u) = P(X^o \leq t, Y^o \leq u)$ by estimating cumulative mark-specific hazard function $\Lambda(t,u)$. For this data setting, we propose an inverse probability weighting (IPW) estimator and an imputation estimator via using the induced binary responses $B_i(t,u) = I(X_i^o \leq t, Y_i^o \leq u)$, i=1, $\cdots$, n.

In a colorectal cancer cohort study, the clinical trial might be terminated before the study endpoint is reached. The terminal events, such as death, preclude any further medical costs and definitely relate with (X^o, Y^o). For instance, it is instinctive to anticipate that poor health patients prone to death might raise larger medical costs. As long as the death arises before the first claim, the medicare reimbursements of the departed would be zero rather than missing. Instead of only considering the joint distrbution of claiming times and reimbursements, it is more meaningful to exploit the joint distribution for those who have medicare reimbursement prior to death, say, $F^*(t, u) = P(X^o \leq t, Y^o \leq u | D > X^o)$, where D denotes the terminal time. In this study, our research focuses on seeking for an appropriate estimation method of $F^*(t, u)$ based on the transformed cost data $\{X_i^*,\ \delta_{X_i^o},\ \delta_{C_i},\ Y_i^*\}_{i=1}^n$, where $X_i^* = X_i^o \wedge C_i \wedge D_i$, $Y_i^* = Y_i^o \delta_{X_i^*}$, $\delta_{X_i^o} = I(X_i^* = X_i^o)$, and $\delta_{C_i} = I(X_i^* = C_i)$. In addition, we provide the estimator for the probability of cost occurring before death $P(D > X^o)$. and the mean cost of patients, which can be derived as $E[Y^o] = \{\int u d_u F^*(t, u)\} P(D > X^o)$.

We divide the rest of this thesis into the following parts. In Chapter 2, we briefly reviews the estimation method of Huang-Louis and propose the IPW and imputation estimators for $F(t, u)$. Moreover, the asymptotic properties and inference procedures are developed in this chapter. In Chapter 3, we further deal with the occurrence of terminal events in estimation. In Chapter 4, a class of simulations to investigate the finite-sample properties of the estimators and the performance of our proposed procedures. We present an application of our methods to the colorectal cancer data retrieved from the Surveillance, Epidemiology, and End Results (SEER) Medicare database in Chapter 5. Finally, a concise discussion and a further research topic are provided in Chapter 6.

Chapter 2

Estimation and Inference Procedures without a Terminal Event

2.1 Review of Huang-Louis Estimator

Baed on the data $\{(X_i,\ \delta_i),\ Y_i\}_{i=1}^n$, Huang and Louis (1998) proposed a non-parametric estimator $\widehat{F}_{HL}(t,u)$ for $F(t,u)$. Before introducing their estimation method, some concise notations are presented first. Let $S_{X^o}(t) = 1-F(t,\infty),\ S_X(t) = P(X > t),\ F_{XY}(t,u) = P(X \le t, Y \le u, \delta = 1),\ S_C(t) = P(C > t),$ and $\Lambda(t,u) = \int_0^t S_{X^o}^{-1}(s) d_s F(s,u)$ be the cumulative cost-specific hazard function, where $\int d_s F(s,u)$ is a Lebesgue-Stieltjes integration over s for fixed u. It was derived that $F(t,u)$ can be expressed as

$$F(t,u) = \int_0^t S_{X^o}(s) d_s \Lambda(s,u) = \int_0^t \prod_{[0,s)} \{1 - d_v \Lambda_{X^o}(v)\} d_s \Lambda(s,u), \qquad (2.1.1)$$

where $\prod$ denotes the product integral and $\Lambda_{X^o}(v) = \Lambda(v,\infty)$.

Under the assumption of random censorship (A1: C is independent of (X^o, Y^o)), one has

$$S_X(t) = S_{X^o}(t)S_C(t) \text{ and } d_tF_{XY}(t,u) = S_C(t)\{d_tF(t,u)\}. \qquad (2.1.2)$$

Substituting the empirical estimators $\widehat{F}_{XY}(t,u) = n^{-1}\sum_{i=1}^{n}\delta_i B_i(t,u)$ and $\widehat{S}_X(t) = n^{-1}\sum_{i=1}^{n} I(X_i > t)$ for $F_{XY}(t,u)$ and $S_X(t)$, an estimator for $\Lambda(t,u)$ was proposed by

$$\widehat{\Lambda}(t,u) = \int_0^t \frac{d_s\widehat{F}_{XY}(s,u)}{\widehat{S}_X(s^-)}. \qquad (2.1.3)$$

It is straightforward to obtain an estimator from (2.1.1) and (2.1.3) as

$$\widehat{F}_{HL}(t,u) = \int_0^t \prod_{[0,s)}\{1 - d_v\widehat{\Lambda}_{X^o}(v)\}d_s\widehat{\Lambda}(s,u), \qquad (2.1.4)$$

with $\widehat{\Lambda}_{X^o}(v) = \widehat{\Lambda}_(v,\infty)$. In the following theorem, we state the uniform consistency and asymptotic Gaussian process of $\widehat{F}_{HL}(t,u)$. Further assumptions (A2: No common jump points in the distributions $S_{X^o}(t)$ and $S_C(t)$) and (A3: $S_C(t)$ and $F(t,u)$ are absolutely continuous on $\Omega = \{(t,u) : 0 < t \leq \tau,\ u > 0\}$ with $\tau = \sup\{t : S_X(t) \geq \epsilon > 0\}$ for some $\epsilon > 0$) are made throught the thesis for the main results.

Theorem 2.1.1. *Supposed that assumptions (A1)-(A3) are satisfied. Then,*

$$\sup_{(t,u)\in\Omega} \mid \widehat{F}_{HL}(t,u) - F(t,u) \mid \xrightarrow{p} 0, \qquad (2.1.5)$$

and $n^{1/2}(\widehat{F}_{HL}(t,u) - F(t,u))$ *converges weakly to a mean zero Gaussian process with variance-covariance function* $\Gamma_1(w_1,w_2) = Cov(Z_1(w_1), Z_1(w_2))$ *for* $w_j = (t_j,u_j) \in \Omega$*, j=1,2, where* $Z_1(t,u) = \int_0^t F(s,u)d_s\varphi(s,\infty) + \int_0^t S_{X^o}(s^-)d_s\varphi(s,u) - F(t,u)\varphi(t,u)$ *and* $\varphi(t,u) = S_X^{-1}(X^-)\delta B(t,u) - \int_0^t S_X^{-2}(s^-)I(X \geq s)d_sF_{XY}(s,u)$.

2.2 IPW Estimator

The IPW technique has been widely adopted in estimation for dealing with the biased sample due to censoring. The main idea is to use subjects with available $B_i(t,u)$'s and weight each observation by the inverse of the selection probability $\pi_t = P(V_t = 1|X^o)$, where $V_t = V_{1t} + V_{2t}$ with $V_{1t} = I(X > t)$ and $V_{2t} = I(X \leq t, \delta = 1)$. Under the validity of random censorship (A1), the selection probability π_t is desired to be $V_{1t}S_C(t) + V_{2t}S_C(X^o)$. By the property,

$$E[V_t \pi_t^{-1}(B(t,u) - F(t,u))] = 0 \qquad (2.2.1)$$

and substituting a consistent estimator $\widehat{S}_C(t)$ for $S_C(t)$, an IPW estimator for $F(t,u)$ can be constructed as

$$\widehat{F}_{IPW}(t,u) = \frac{\sum_{i=1}^{n} \delta_i \widehat{S}_C^{-1}(X_i^o) B_i(t,u)}{\sum_{i=1}^{n} V_{it}\widehat{\pi}_{it}^{-1}}, \qquad (2.2.2)$$

where $\widehat{\pi}_{it} = V_{i1t}\widehat{S}_C(t) + V_{i2t}\widehat{S}_C(X_i^o)$. Naturally, the Kaplan-Meier estimator is used to estimate $S_C(t)$. The asymptotic Gaussian process of $n^{1/2}(\widehat{F}_{IPW}(t,u) - F(t,u))$ is established below.

Theorem 2.2.1. *Suppose that assumptions (A1)-(A3) hold. Then,* $n^{1/2}(\widehat{F}_{IPW}(t,u)$ $- F(t,u))$ *converges weakly to a Gaussian process with mean zero and variance-covariance function* $\Gamma_2(w_1, w_2) = Cov(Z_{i2}(w_1), Z_{i2}(w_2))$*, where* $Z_{i2}(t,u) = E[\psi_{ij}(t,u) + \psi_{ji}(t,u)|((X_i, \delta_i), Y_i)]$ *with* $M_{C_i}(t) = I(X_i \leq t)(1 - \delta_i) + \int_0^t I(X_i \geq v)d(lnS_C(v))$ *and*

$$\psi_{ij}(t,u) = \{\frac{V_{i1t}}{S_C(t)}(1 + \frac{1}{n}\int_0^t \frac{dM_{C_j}(s)}{S_X(s^-)}) + \frac{V_{i2t}}{S_C(X_i^o)}(1 + \frac{1}{n}\int_0^{X_i^o} \frac{dM_{C_j}(s)}{S_X(s^-)})\}(B_i(t,u) - F(t,u)).$$

Proof. From (2.2.2), one has

$$n^{1/2}(\widehat{F}_{IPW}(t,u) - F(t,u)) = \frac{n^{-1/2}\sum_{i=1}^{n} V_{it}\widehat{\pi}_{it}^{-1}(B_i(t,u) - F(t,u))}{n^{-1}\sum_{i=1}^{n} V_{it}\widehat{\pi}_{it}^{-1}}. \tag{2.2.3}$$

By the boundness of V_{it}'s, the uniform convergence of $\widehat{S}_C(t)$, and the Euclidean class of $\{V_{it}\pi_{it}^{-1} : 0 < t \leq \tau\}$ (cf. Akrita (1994), Pakes and Pollard (1989), and Pollard (1990)), it is entailed that $n^{-1}\sum_{i=1}^{n} V_{it}\pi_{it}^{-1}$ uniformly converges to one. Thus,

$$n^{1/2}(\widehat{F}_{IPW}(t,u) - F(t,u)) = n^{-1/2}\sum_{i=1}^{n} V_{it}\widehat{\pi}_{it}^{-1}(B_i(t,u) - F(t,u)) + r_{1n}(t,u), \tag{2.2.4}$$

where $\sup_{(t,u)\in\Omega} |r_{1n}(t,u)| = o_p(1)$. The first order Taylor expansion of the dominating term in (2.2.4) with respect to $\widehat{S}_C(t) = S_C(t)$ yields that

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n} V_{it}\widehat{\pi}_{it}^{-1}(B_i(t,u) - F(t,u)) &= n^{-1/2}\sum_{i=1}^{n} V_{it}\pi_{it}^{-1}(B_i(t,u) - F(t,u)) - \\ n^{-1/2}\sum_{i=1}^{n}\{\frac{V_{i1t}}{S_C(t)}(\frac{\hat{S}_C(t)}{S_C(t)} - 1) + \frac{V_{i2t}}{S_C(X_i^o)}&(\frac{\hat{S}_C(X_i^o)}{S_C(X_i^o)} - 1)\}(B_i(t,u) - F(t,u)) + r_{2n}(t), \end{aligned} \tag{2.2.5}$$

where $\sup_{t\in(0,\tau]} |r_{2n}(t)| = o_p(n^{-1/2})$. Since $\{\widehat{S}_C(t)/S_C(t) - 1\}$ can be uniformly approximated by $\{-n^{-1}\sum_{i=1}^{n}\int_0^t dM_{C_i}(u)/S_X(u^-)\}$ (cf. Fleming and Harrington (1991)). It follows from (2.2.4)-(2.2.5) that

$$n^{1/2}(\widehat{F}_{IPW}(t,u) - F(t,u)) = \{n^{1/2}(n-1)\}^{-1}\sum\sum_{i\neq j}\psi_{ij}(t,u) + r_{3n}(t,u), \tag{2.2.6}$$

where $\sup_{(t,u)\in\Omega} |r_{3n}(t,u)| = o_p(n^{-1/2})$. By the Euclidean class of $\{\psi_{ij}(t,u) : (t,u) \in \Omega\}$, the decomposition of a U-statistic into the sum of degenerate U-statistics, and Corollary 4 of Sherman (1994), $n^{1/2}(\widehat{F}_{IPW}(t,u) - F(t,u))$ can be uniformly approximated by the term of independent and identically distributed random quantities

$n^{-1/2}\sum_{i=1}^{n} Z_{i2}(t,u)$. By the functional central limit theorem (Pollard (1990)) and the uniform convergence of $n^{-1}\sum_{i=1}^{n}\{V_{it}/\pi_{it}\}$ to one, the proof for Theorem 2.2.1 is completed. □

2.3 Imputation Estimator

Applying the method of Buckley and James (1979), we propose a alternative estimator based on the considered data setting. In this estimation method, the unavailable statuses $B_i(t,u)$'s are substituted by the corresponding expectations $E[B_i(t,u)|X_i,\delta_i=0]$'s. Let $B_i^*(t,u)=I(V_{it}=1)B_i(t,u)+I(V_{it}=0)E[B_i(t,u)|X_i,\delta_i=0]$. A direct calculation ensures that

$$E[B_i^*(t,u)|X_i,\delta_i]=\sum_{k=0}^{1}E[B_i^*(t,u)|X_i,\delta_i]I(V_{it}=k)$$
$$=\sum_{k=0}^{1}E[B_i(t,u)|X_i,\delta_i]I(V_{it}=k)=E[B_i(t,u)|X_i,\delta_i], \tag{2.3.1}$$

which implies that $E[B_i^*(t,u)]=E[B_i(t,u)]=F(t,u)$. Under assumption (A1), we further derive that

$$E[B_i(t,u)|X_i=x,\delta_i=0]=S_{X^o}^{-1}(x)\{F(t,u)-F(x,u)\}. \tag{2.3.2}$$

Following from (2.3.1)-(2.3.2), an estimation procedure for $F(t,u)$ is proposed. The imputation estimator $\widehat{F}_{IM}(t,u)$ is obtained via solving

$$n^{-1}\sum_{i=1}^{n}(B_i^*(t,u)-F(t,u))$$
$$=n^{-1}\sum_{i=1}^{n}\{(B_i(t,u)-F(t,u))I(V_{it}=1)+\frac{(1-\widehat{S}_{X^o}(X_i))F(t,u)-F(X_i,u)}{\widehat{S}_{X^o}(X_i)}I(V_{it}=0)\}$$
$$\triangleq 0, \tag{2.3.3}$$

where $\widehat{S}_{X^o}(t)$ is the Kaplan-Meier estimator of $S_{X^o}(t)$.

Note that $F(X_i, u)$ in (2.3.3) is unknown when $V_{it} = 0$. Generally, a consistent estimator $\widehat{F}(X_i, u)$ is substituted for $F(X_i, u)$. Let A_t be the collection of ordered censoring times $C_{(1)} < C_{(2)} < \cdots < C_{(k_t)} \leq t$ with size k_t. The joint distribution $F(C_{(1)}, u)$ is first estimated by

$$\widehat{F}_{IM}(C_{(1)}, u) = n^{-1} \sum_{i=1}^{n} B_i(C_{(1)}, u). \tag{2.3.4}$$

Subsequently, the estimator for $F(C_{(j)}, u)$, j=2,$\cdots$, k_t, can be obtained as

$$\widehat{F}_{IM}(C_{(j)}, u) = \frac{\sum_{i=1}^{n} B_i(C_{(j)}, u) I(V_{iC_{(j)}} = 1) - \sum_{l=1}^{j-1} \widehat{S}_{X^o}^{-1}(C_{(l)}) \widehat{F}(C_{(l)}, u)}{\sum_{i=1}^{n} I(V_{iC_{(j)}} = 1) - \sum_{l=1}^{j-1} \widehat{S}_{X^o}^{-1}(C_{(l)})(1 - \widehat{S}_{X^o}(C_{(l)}))}. \tag{2.3.5}$$

An explicit expression for the estimator of $F(t, u)$ is then derived to be

$$\widehat{F}_{IM}(t, u) = \frac{\sum_{i=1}^{n} B_i(t, u) I(V_{it} = 1) - \sum_{i=1}^{n} \widehat{S}_{X^o}^{-1}(X_i) \widehat{F}(X_i, u) I(V_{it} = 0)}{\sum_{i=1}^{n} I(V_{it} = 1) - \sum_{i=1}^{n} \widehat{S}_{X^o}^{-1}(X_i)(1 - \widehat{S}_{X^o}(X_i)) I(V_{it} = 0)}. \tag{2.3.6}$$

Remark: Since the number of censoring times prior to t is random and might be considerably large, it becomes cumbersome to develop the inference procedure for $F(t, u)$ based on $\widehat{F}_{IM}(t, u)$. Currently, there is still no existing statistical methodology to facilitate this obstacle.

2.4 Inferences Procedures

In this subsection, pointwise and simultaneous confidence bands for $F(t, u)$ are constructed based on the asymptotic Gaussian processes of $\widehat{F}_{HL}(t, u)$ and $\widehat{F}_{IPW}(t, u)$. As shown in Theorem 2.1.1, the limiting process $n^{1/2}(\widehat{F}_{HL}(t, u) - F(t, u))$ is uniformly asymptotically equivalent to $n^{-1/2} \sum_{i=1}^{n} Z_{i1}(t, u)$. The variance-covariance function of $\widehat{F}_{HL}(t,u)$ is suggested to be estimated by $\widehat{\Gamma}_1(w_1,w_2) = n^{-1}\sum_{i=1}^{n}\widehat{Z}_{i1}(w_1)\widehat{Z}_{i1}(w_2)$, where

$$\widehat{Z}_{i1}(t,u) = \int_0^t \widehat{F}_{HL}(s,u)ds\widehat{\varphi}_i(s) + \int_0^t \widehat{S}_{X^o}(s)ds\widehat{\varphi}_i(s,u) - \widehat{F}_{HL}(t,u)\widehat{\varphi}_i(t,u) \quad (2.4.1)$$

with $\widehat{\varphi}_i(t,u) = \widehat{S}_X^{-1}(X_i)\delta_i B_i(t,u) - \int_0^t \widehat{S}_X^{-2}(s)I(X_i \geq s)ds\widehat{F}_{XY}(s,u)$. Similarly, the limiting process of $n^{1/2}(\widehat{F}_{IPW}(t,u) - F(t,u))$ is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n} Z_{i2}$ (t,u). Let $\widehat{M}_{C_j}(t) = I(X_j \leq t)(1-\delta_j) + \int_0^t I(X_j \geq v)d(\ln\widehat{S}_C(v))$. An estimator for the variance-covariance function $\Gamma_2(w_1,w_2)$ is proposed to be

$$\widehat{\Gamma}_2(w_1,w_2) = n^{-1}\sum_{i=1}^{n}\widehat{Z}_{i2}(w_1)\widehat{Z}_{i2}(w_2), \quad (2.4.2)$$

where $\widehat{Z}_{i2}(t,u) = (n-1)^{-1}\sum_{j\neq i}(\widehat{\psi}_{ij}(t,u) + \widehat{\psi}_{ji}(t,u))$ and

$$\widehat{\psi}_{ij}(t,u) = (\frac{1}{n}\frac{V_{i1t}}{\widehat{S}_C(t)}\int_0^t \frac{d\widehat{M}_{C_j}(s)}{\widehat{S}_X(s^-)} + \frac{1}{n}\frac{V_{i2t}}{\widehat{S}_C(X_i^o)}\int_0^{X_i^o}\frac{d\widehat{M}_{C_j}(s)}{\widehat{S}_X(s^-)})(B_i(t,u) - \widehat{F}_{IPW}(t,u)).$$

The uniform consistency of $\widehat{\Gamma}_1(w_1,w_2)$ and $\widehat{\Gamma}_2(w_1,w_2)$ are established in the following theorem.

Theorem 2.4.1. *Supposed that assumptions (A1)-(A3) hold. Then,*

$$\sup_{w_1,w_2\in\Omega} \mid \widehat{\Gamma}_l(w_1,w_2) - \Gamma_l(w_1,w_2) \mid \xrightarrow{p} 0, \ \text{as } n \longrightarrow \infty, l = 1,2. \quad (2.4.3)$$

Proof. By the uniform convergence of empirical estimators (Pollard (1990)), we can show that $\widehat{F}_{XY}(t,u)$ and $\widehat{S}_X(t)$ converge to $F_{XY}(t,u)$ and $S_X(t)$ uniformly in (t,u). Together with (2.4.1), Theorem 2.1.1, and the boundness of $\delta_i B_i(t,u)$'s, it is ensured that

$$\sup_{(t,u)\in\Omega} \mid \widehat{\varphi}_i(t,u) - \varphi_i(t,u) \mid \xrightarrow{p} 0 \text{ as } n \longrightarrow \infty, i = 1,\cdots,n. \quad (2.4.4)$$

One further derives by the intergration by parts that

$$\sup_{(t,u)\in\Omega} \mid \widehat{Z}_{i1}(t,u) - Z_{i1}(t,u) \mid \xrightarrow{p} 0 \text{ as } n \longrightarrow \infty, i = 1, \cdots, n. \tag{2.4.5}$$

From (2.4.8), Theorem 2.2.1, and the inequality

$$\mid \widehat{\Gamma}_1(w_1,w_2) - \Gamma_1(w_1,w_2) \mid \leq \mid \widehat{\Gamma}_1(w_1,w_2) - n^{-1}\sum_{i=1}^{n} Z_{i1}(w_1)Z_{i1}(w_2) \mid$$

$$+ \mid n^{-1}\sum_{i=1}^{n} Z_{i1}(w_1)Z_{i1}(w_2) - \Gamma_1(w_1,w_2) \mid, \tag{2.4.6}$$

the uniform consistency of $\widehat{\Gamma}_1(w_1,w_2)$ is obtained.

For the uniform consistency of $\widehat{\Gamma}_2(w_1,w_2)$, it is implied from (2.4.2) and Theorem 2.2.1 that

$$\widehat{\Gamma}_2(w_1,w_2) = n^{-1}\sum_{i=1}^{n} \widehat{Z}_{i2}(w_1)\widehat{Z}_{i2}(w_2)$$

$$= \{n(n-1)^2\}^{-1}\sum_{i,j,k} (\widehat{\psi}_{ij}(w_1) + \widehat{\psi}_{ji}(w_1))(\widehat{\psi}_{ik}(w_2) + \widehat{\psi}_{ki}(w_2)). \tag{2.4.7}$$

The expression of $\widehat{\Gamma}_2(w_1,w_2)$ in (2.4.7) and the uniform convergence of $\widehat{S}_X(t)$, $\widehat{S}_C(t)$, and $\widehat{F}_{IPW}(t,u)$ entail that $\widehat{\Gamma}_2(w_1,w_2)$ is uniformly approximated by

$$\{n(n-1)(n-2)\}\sum_{i\neq j\neq k} (\psi_{ij}(w_1) + \psi_{ji}(w_1))(\psi_{ik}(w_2) + \psi_{ki}(w_2)). \tag{2.4.8}$$

Since $\{\psi_{ij}(t,u) : (t,u) \in \Omega\}$ is Euclidean, the term in (2.4.8) converges to $E[(\psi_{ij}(w_1) + \psi_{ji}(w_1))(\psi_{ik}(w_2) + \psi_{ki}(w_2))] = E[Z_{i2}(w_1)Z_{i2}(w_2)]$ uniformly in (w_1,w_2). □

The limiting Gaussian processes in Theorems 2.1.1 and 2.2.1 as well as the estimated variance-covariance functions enable us to construct an approximated $(1-\alpha)$ pointwise confidence interval for $F(t,u)$ via either

$$\widehat{F}_{HL}(w) \pm z_{1-\alpha/2}\widehat{\Gamma}_1^{1/2}(w,w) \text{ or } \widehat{F}_{IPW}(w) \pm z_{1-\alpha/2}\widehat{\Gamma}_2^{1/2}(w,w), \tag{2.4.3}$$

where $z_{1-\alpha}$ is the $100(1-\alpha)th$ percentile of a univariate standard normal distribution and $w = (t, u)$. To draw inference on the pattern of $F(t, u)$, a simultaneous confidence band can also be established based on the quantities:

$$\widehat{K}_l = \sup_{w \in \Upsilon} \mid \frac{n^{-1/2} \sum_{i=1}^{n} M_i \widehat{Z}_{il}(w)}{\widehat{\Gamma}_l^{1/2}(w)} \mid, l = 1, 2, \tag{2.4.4}$$

where Υ is a region of interest within Ω, and $\{M_i : i = 1, \cdots, n\}$ are independent realizations of standard normal variable and are independent of $\{(X_i, \delta_i), Y_i\}_{i=1}^{n}$. An approximated $(1-\alpha)$ simultaneous confidence band for $F(t, u)$ is then constructed via either

$$\widehat{F}_{HL}(w) \pm n^{-1/2} Q_{1-\alpha}(\widehat{K}_1) \widehat{\Gamma}_1^{1/2}(w, w) \text{ or } \widehat{F}_{IPW}(w) \pm n^{-1/2} Q_{1-\alpha}(\widehat{K}_2) \widehat{\Gamma}_2^{1/2}(w, w), \tag{2.4.5}$$

where $Q_{1-\alpha}(\widehat{K}_l)$ is the $100(1-\alpha)th$ percentile of realizations of (2.4.4) computed based on a large number of generations of $\{M_i ; i = 1, \cdots, n\}$, $l = 1, 2$.

Chapter 3

Estimation and Inference Procedures with Terminal Events

By adapting the IPW method mentioned in Chapter 2 for the appearance of terminal events, we use subjects with available $B_i(t,u)'s$ and weight each subject with $\pi_i = P(\delta_{X_i^o} = 1|X_i^o)'s$. Under the assumption of random censorship (AA1: C is independent of (X^o, Y^o, D)) and assumption (AA2: The distribution functions of X^o, C and D do not have jump points in common), π_i is derived to be $S_C(X_i^o)$. Using

$$E[\delta_{X^o} S_C^{-1}(X^o)(B(t,u) - F^*(t,u))] = 0 \tag{3.1.1}$$

and replacing $S_C(t)$ by a consistent estimator $\widehat{S}_C(t)$, $F^*(t,u)$ is proposed to be estimated by

$$\widehat{F}^*(t,u) = \frac{\sum_{i=1}^n \delta_{X_i^o} \widehat{S}_C^{-1}(X_i^o) B_i(t,u)}{\sum_{i=1}^n \delta_{X_i^o} \widehat{S}_C^{-1}(X_i^o)}. \tag{3.1.2}$$

Generally, it is reasonably to estimate $S_C(t)$ by the Kaplan-Meier estimator. Based on (3.1.2), the probability $P(D > X^o)$ and the mean cost $E[Y^o]$ are naturally estimated by

$$\widehat{P}(D > X^o) = n^{-1}\sum_{i=1}^{n} \delta_{X_i^o}\widehat{S}_C^{-1}(X_i) \text{ and } \widehat{E}[Y^o] = n^{-1}\sum_{i=1}^{n} \delta_{X_i^o}\widehat{S}_C^{-1}(X_i^o)Y_i^o. \qquad (3.1.3)$$

To establish the asymptotic properties of $n^{1/2}(\widehat{F}^*(t,u) - F^*(t,u))$, the condition (AA3: $S_C(t)$, $S_D(t)$, and $F^*(t,u)$ are absolutely continuous on $(0,\tau^*]$ and $\Omega^* = \{(t,u) : 0 < t \le \tau^*,\ u > 0\ \}$ with $\tau^* = \sup\{t : S_{X^*}(t) \ge \epsilon > 0\}$ for some $\epsilon > 0$.) is further made in the following theorem.

Theorem 3.1. Suppose that assumptions (AA1)-(AA3) are satisfied. $n^{1/2}(\widehat{F}^*(t,u) - F^*(t,u))$ converges weakly to a mean-zero Gaussian process with variance-covariance function $\Gamma^*(w_1,w_2) = Cov(Z_i^*(w_1), Z_i^*(w_2))$ for $w_1, w_2 \in \Omega^*$, where $Z_i^*(t,u) = E[\psi_{ij}^*(t,$

$u) + \psi_{ji}^*(t,u)|((X_i^*, \delta_{X_i^o}, \delta_{C_i}), Y_i^*)]$, $M_{C_i}^*(t) = \delta_{C_i}I(X_i^* \le t) + \int_0^t I(X_i^* \ge v)d(\ln S_C(v))$ and

$$\psi_{ij}^*(t,u) = \frac{\delta_{X_i^o}}{\{S_C(X_i^o)P(D > X^o)\}}\left(1 + \frac{1}{n}\int_0^{X_i^o} S_{X^*}^{-1}(s^-)dM_{C_j}^*(s)\right)(B_i(t,u) - F^*(t,u)).$$

Proof. From (3.1.2), $n^{1/2}(\widehat{F}^*(t,u) - F^*(t,u))$ can be re-expressed as

$$\frac{n^{-1/2}\sum_{i=1}^n \delta_{X_i^o}\widehat{S}_C^{-1}(X_i^o)(B_i(t,u) - F^*(t,u))}{n^{-1}\sum_{i=1}^n \delta_{X_i^o}\widehat{S}_C^{-1}(X_i^o)}. \qquad (3.1.4)$$

By the boundedness of $\delta_{X_i^o}$'s and the uniform convergence of $\widehat{S}_C(t)$, the denominator term in (3.1.4) is shown to be asymptotically equivalent to $P(D > X^o)$. Thus,

$$n^{-1/2}(\widehat{F}^*(t,u) - F^*(t,u)) = \frac{n^{-1/2}\sum_{i=1}^n \delta_{X_i^o}\widehat{S}_C^{-1}(X_i^o)(B_i(t,u) - F^*(t,u))}{P(D > X^o)} + r_{1n}^*(t,u), \qquad (3.1.5)$$

where $\sup\limits_{(t,u)\in\Omega^*} |r^*_{1n}(t,u)| = o_p(1)$. Applying the Taylor expansion theorem to the numerator term in (3.1.5) with respect to $\widehat{S}_C(t) = S_C(t)$, it can be derived to be

$$n^{-1/2}\sum_{i=1}^{n}\frac{\delta_{X^o_i}}{\widehat{S}_C(X^o_i)}(B_i(t,u) - F^*(t,u)) = n^{-1/2}\sum_{i=1}^{n}\frac{\delta_{X^o_i}}{S_C(X^o_i)}(B_i(t,u) - F^*(t,u)) -$$

$$n^{-1/2}\sum_{i=1}^{n}\{\frac{\delta_{X^o_i}}{S_C(X^o_i)}(\frac{\hat{S}_C(X^o_i)}{S_C(X^o_i)} - 1)(B_i(t,u) - F^*(t,u))\} + r^*_{2n}(t), \qquad (3.1.6)$$

where $\sup\limits_{t\in(0,\tau^*]} |r^*_{2n}(t)| = o_p(n^{-1/2})$. Together with (3.1.5) and the proof argument for Theorem 2.2.1, we obtain that

$$n^{1/2}(\widehat{F}^*(t,u) - F^*(t,u)) = \{n^{1/2}(n-1)\}^{-1}\sum\sum_{i\neq j}\psi^*_{ij}(t,u) + r^*_n(t), \qquad (3.1.7)$$

where $\sup\limits_{t\in(0,\tau^*]} |r^*_n(t)| = o_p(n^{-1/2})$. The asymptotic Gaussian process of $\widehat{F}^*(t,u)$ is developed. □

Let $\widehat{M}^*_{C_i}(t) = \delta_{C_i} I(X^*_i \leq t) + \int_0^t I(X^*_i \geq v) d(\ln\widehat{S}_C(v))$, $i = 1, \cdots, n$, and $\widehat{S}_{X^*}(t) = n^{-1}\sum_{i=1}^n I(X^*_i > t)$. The variance-covariance function $\Gamma^*(w_1, w_2)$ is suggested to be estimated by $\widehat{\Gamma}^*(w_1, w_2) = n^{-1}\sum_{i=1}^n \widehat{Z}^*_i(w_1)\widehat{Z}^*_i(w_2)$ with $\widehat{Z}^*_i(t,u) =$
$(n-1)^{-1}\sum\limits_{j\neq i}(\widehat{\psi}^*_{ij}(t,u)$
$+ \widehat{\psi}^*_{ji}(t,u))$ and

$$\widehat{\psi}^*_{ij}(t,u) = \delta_{X^o_i}\{\widehat{S}_C(X^o_i)\widehat{P}(D > X^o)\}^{-1}\frac{1}{n}\int_0^{X^o_i}\frac{d\widehat{M}^*_{C_j}(s)}{\widehat{S}_{X^*}(s^-)}(B_i(t,u) - \widehat{F}^*(t,u)). \qquad (3.1.8)$$

The uniform consistency of $\widehat{\Gamma}^*(w_1, w_2)$ is given in the following theorem.

Theorem 3.2. Supposed that assumptions (AA1)-(AA3) hold. Then,

$$\sup_{w_1,w_2\in\Omega^*} \mid \widehat{\Gamma}^*_1(w_1,w_2) - \Gamma^*_1(w_1,w_2) \mid \xrightarrow{p} 0, \text{ as } n \longrightarrow \infty. \qquad (3.1.9)$$

Proof. By using the techniques in the proof of Theorem 3.1, the uniform convergence of $\widehat{P}(D > X^o)$ is obtained. Paralleling the argument for the uniform consistency of $\widehat{\Gamma}_2(w_1, w_2)$ in Theorem 2.4.1, the uniform consistency of $\widehat{\Gamma}^*(w_1, w_2)$ is developed. $\square$

Similarly to the aforementioned procedure, approximated $(1-\alpha)$ pointwise and simultaneous confidence intervals are seperately constructed via

$$\widehat{F}^*(w) \pm z_{1-\alpha/2}\widehat{\Gamma}^{*1/2}(w,w) \text{ and } \widehat{F}^*(w) \pm n^{-1/2}Q_{1-\alpha}(\widehat{K}^*)\widehat{\Gamma}^{*1/2}(w,w), \qquad (3.1.10)$$

where

$$\widehat{K}^* = \sup_{w \in \Upsilon^*} \mid \frac{n^{-1/2}\sum_{i=1}^{n} M_i^* \widehat{Z}_i^*(w)}{\widehat{\Gamma}^{*1/2}(w)} \mid$$

with Υ^* is a region of interest within Ω^*, and $(M_i^* : i = 1, \cdots, n)$ are independent realizations of standard normal variable and are independent of $\{(X_i^*, \delta_{X_i^o}, \delta_{C_i}), Y_i^*\}_{i=1}^n$.

Chapter 4

Numerical Studies

In this chapter, we conduct two simulation scenarios to investigate the finite sample properties of proposed estimators and the performance of inference procedures. One is for the case of censoring data and the other accommodates the appearance of terminal events. In the simulation process, data are repeatedly generated 500 times with the sample sizes of 200 and 400, and the censoring rates of 30% and 50%. The estimators are evaluated at the selected grid points (t, u), where u takes values of 0.25, 0.5, 0.75, and 1, and t takes values of 0.2231, 0.5108, 0.9163, and 1.6094 with the cumulative probabilities of X^o being 0.2, 0.4, 0.6, and 0.8.

4.1 Simulation Setting of (X^o, Y^o)

The pair random vector (X^o, Y^o) is specified from the Frank's bivariate family (Genest (1987)), in which

$$F^{(\alpha)}_{X^oY^o}(t,u) = \begin{cases} \log_\alpha\{1 + (\alpha^{F_{X^o}(t)} - 1)(\alpha^{F_{Y^o}(u)} - 1)/(\alpha - 1)\}, & \alpha \neq 1 \\ F_{X^o}(t)F_{Y^o}(u), & \alpha = 1. \end{cases}$$

Simulations are implemented with $\alpha = exp(-10)$, which implies a positive association between the claiming time and the medical cost.

4.2 Senario I - Without a terminal event

In the section, we examine the finite sample properties of $\widehat{F}_{HL}(t, u)$, $\widehat{F}_{IPW}(t, u)$, and $\widehat{F}_{IM}(t, u)$, and evaluate the inference procedures based on $\widehat{F}_{HL}(t, u)$ and $\widehat{F}_{IPW}(t, u)$. The censoring time C is independently generated from an exponential distribution with different scale parameters 0.5 and 1 for the expected censoring rates of 30% and 50%.

Tables 4.1-4.6 exhibit the averages and standard deviations of 500 estimates, and the averages of 500 standard errors based on (2.1.4), (2.2.2), and (2.3.7) at the selected points. We detect that the averages of 500 estimates $\widehat{F}_{IPW}(t, u)$ are more close to $F(t, u)$ than those of $\widehat{F}_{HL}(t, u)$, especially for a higher censoring rate. Furthermore, $\widehat{F}_{IM}(t, u)$ is found to have larger biases at points of (1.6094, 0.75) and (1.6094, 1.0) when the sample size is small and the censoring rate is high. The biases of these estimators are negligible when the sample size is large enough. The standard deviations of three estimates are almost the same. As expected, the standard deviations decrease with increasing sample size and decreasing censoring rate. It is revealed in these tables that the averages of 500 standard errors of $\widehat{F}_{HL}(t, u)$ and $\widehat{F}_{IPW}(t, u)$ are very close to the standard deviations of their 500 estimates. Note that the averages of 500 standard errors of $\widehat{F}_{HL}(t, u)$ diverge from the standard deviations of estimates as the value of time is large, while those of $\widehat{F}_{IPW}(t, u)$ seem to be relatively accurate. In tables 4.7-4.8, 0.95 pointwise confidence intervals and the corresponding empirical coverage probabilities are provided. Generally, the coverage probabilities of 0.95 pointwise confidence intervals based on (2.2.2) are fairly close to the nominal level. However, the empirical coverage probability of confidence intervals based on (2.1.4) are much higher than the expected nominal level.

4.3 Senario II - With terminal events

In this simulation study, the finite sample properties of $\widehat{F}^*_{IPW}(t,u)$ and the performance of inference procedure are investigated as terminal events arise. For the design of mixture rates of censoring and death, C is independently generated from an exponential distribution with parameter a and the terminal time D is designed to follow an exponential distribution with rate $bI(X^0 \leq 1, Y^0 \leq 0.5) + b$. The parameters (a,b) are set to be $(0.5, 0.01)$ and $(0.6, 0.3)$ so that the mixture rates of 30% and 50% are achieved in the simulated data.

Tables 4.9-4.10 display the averages and standard deviations of 500 estimates, and the averages of 500 standard errors based on (3.1.2) at the considered points. The averages of 500 estimates generally close to the true values of $F^*(t,u)$'s. The biases are apparently reduced when the sample size is large or the mixture rate is small. Moreover, the variation of estimator will decrease and the accuracy of estimated variances will be improved as the sample size increases or the mixture rate decreases. Table 4.11 exhibits the empirical coverage probabilities of 0.95 pointwise confidence intervals for $F^*(t,u)$. The probabilities are generally around the nominal level of 0.95. It is revealed in these tables that the closeness of empirical coverage probabilities to the nominal level relies on the sample size and the censoring rate.

Table 4.1: The averages and the standard deviations (SD) of 500 estimates $\widehat{F}_{HL}(t,u)$ and the averages of 500 standard errors (SE) at the selected points with the sample sizes (n) of 200 and 400, and the censoring rate of 30%

n			200			400		
t	u	$F(t,u)$	Mean	SD	SE	Mean	SD	SE
0.2231	0.25	0.158	0.156	0.0279	0.0269	0.156	0.0182	0.0190
0.5108	0.25	0.231	0.234	0.0320	0.0345	0.232	0.0227	0.0237
0.9163	0.25	0.247	0.249	0.0328	0.0395	0.248	0.0237	0.0278
1.6094	0.25	0.250	0.251	0.0331	0.0564	0.250	0.0238	0.0394
0.2231	0.50	0.196	0.194	0.0307	0.0295	0.193	0.0195	0.0209
0.5108	0.50	0.369	0.371	0.0374	0.0397	0.369	0.0262	0.0282
0.9163	0.50	0.469	0.469	0.0379	0.0512	0.470	0.0278	0.0362
1.6094	0.50	0.496	0.496	0.0385	0.0914	0.497	0.0281	0.0638
0.2231	0.75	0.200	0.197	0.0310	0.0297	0.197	0.0198	0.0211
0.5108	0.75	0.397	0.399	0.0374	0.0404	0.398	0.0261	0.0286
0.9163	0.75	0.581	0.579	0.0376	0.0477	0.580	0.0273	0.0338
1.6094	0.75	0.708	0.703	0.0400	0.0865	0.708	0.0280	0.0607
0.2231	1.0	0.200	0.198	0.0311	0.0298	0.197	0.0198	0.0211
0.5108	1.0	0.400	0.402	0.0376	0.0405	0.400	0.0263	0.0287
0.9163	1.0	0.600	0.598	0.0374	0.0461	0.598	0.0277	0.0327
1.6094	1.0	0.800	0.795	0.0357	0.0458	0.796	0.0265	0.0324

Table 4.2: The averages and the standard deviations (SD) of 500 estimates $\widehat{F}_{IPW}(t,u)$ and the averages of 500 standard errors (SE) at the selected points with the sample sizes (n) of 200 and 400, and the censoring rate of 30%

n			200			400		
t	u	$F(t,u)$	Mean	SD	SE	Mean	SD	SE
0.2231	0.25	0.158	0.158	0.0279	0.0265	0.158	0.0183	0.0187
0.5108	0.25	0.231	0.234	0.0321	0.0318	0.234	0.0227	0.0224
0.9163	0.25	0.247	0.250	0.0329	0.0334	0.250	0.0238	0.0236
1.6094	0.25	0.250	0.252	0.0332	0.0345	0.252	0.0239	0.0243
0.2231	0.50	0.196	0.196	0.0310	0.0289	0.196	0.0195	0.0204
0.5108	0.50	0.369	0.373	0.0375	0.0371	0.371	0.0263	0.0262
0.9163	0.50	0.469	0.471	0.0379	0.0409	0.471	0.0279	0.0289
1.6094	0.50	0.496	0.497	0.0386	0.0440	0.498	0.0281	0.0311
0.2231	0.75	0.200	0.200	0.0313	0.0292	0.200	0.0197	0.0206
0.5108	0.75	0.397	0.401	0.0374	0.0377	0.399	0.0262	0.0267
0.9163	0.75	0.581	0.583	0.0373	0.0416	0.584	0.0275	0.0294
1.6094	0.75	0.708	0.705	0.0400	0.0441	0.711	0.0279	0.0311
0.2231	1.0	0.200	0.200	0.0315	0.0292	0.200	0.0198	0.0206
0.5108	1.0	0.400	0.403	0.0377	0.0378	0.402	0.0264	0.0267
0.9163	1.0	0.600	0.602	0.0372	0.0415	0.602	0.0278	0.0294
1.6094	1.0	0.800	0.799	0.0358	0.0406	0.800	0.0264	0.0287

Table 4.3: The averages and standard deviations (SD) of 500 estimates $\widehat{F}_{IM}(t,u)$ at the selected points with the sample sizes (n) of 200 and 400, and the censoring rate of 30%

n			200		400	
t	u	$F(t,u)$	Mean	SD	Mean	SD
0.2231	0.25	0.158	0.158	0.0279	0.158	0.0183
0.5108	0.25	0.231	0.234	0.0321	0.232	0.0227
0.9163	0.25	0.247	0.249	0.0328	0.248	0.0238
1.6094	0.25	0.250	0.251	0.0330	0.250	0.0238
0.2231	0.50	0.196	0.196	0.0310	0.196	0.0195
0.5108	0.50	0.369	0.372	0.0374	0.370	0.0263
0.9163	0.50	0.469	0.470	0.0378	0.471	0.0278
1.6094	0.50	0.496	0.495	0.0384	0.497	0.0280
0.2231	0.75	0.200	0.200	0.0313	0.200	0.0197
0.5108	0.75	0.397	0.400	0.0374	0.399	0.0262
0.9163	0.75	0.581	0.582	0.0373	0.583	0.0274
1.6094	0.75	0.708	0.702	0.0399	0.709	0.0279
0.2231	1.0	0.200	0.200	0.0314	0.200	0.0198
0.5108	1.0	0.400	0.403	0.0376	0.402	0.0264
0.9163	1.0	0.600	0.601	0.0372	0.602	0.0278
1.6094	1.0	0.800	0.795	0.0356	0.798	0.0264

Table 4.4: The averages and standard deviations (SD) of 500 estimates $\widehat{F}_{HL}(t,u)$ and the averages of 500 standard errors (SE) at the selected points with the sample sizes (n) of 200 and 400, and the censoring rate of 50%

n			200			400		
t	u	$F(t,u)$	Mean	SD	SE	Mean	SD	SE
0.2231	0.25	0.158	0.154	0.0263	0.0273	0.156	0.0192	0.0196
0.5108	0.25	0.231	0.230	0.0323	0.0349	0.230	0.0228	0.0248
0.9163	0.25	0.247	0.246	0.0344	0.0427	0.246	0.0244	0.0304
1.6094	0.25	0.250	0.248	0.0348	0.0696	0.249	0.0248	0.0488
0.2231	0.5	0.196	0.192	0.0298	0.0302	0.192	0.0210	0.0215
0.5108	0.5	0.369	0.367	0.0398	0.0417	0.366	0.0271	0.0296
0.9163	0.5	0.469	0.466	0.0431	0.0565	0.467	0.0306	0.0400
1.6094	0.5	0.496	0.494	0.0456	0.1181	0.494	0.0324	0.0821
0.2231	0.75	0.200	0.196	0.0298	0.0304	0.196	0.0212	0.0217
0.5108	0.75	0.397	0.395	0.0417	0.0423	0.394	0.0272	0.0300
0.9163	0.75	0.581	0.577	0.0457	0.0518	0.576	0.0307	0.0369
1.6094	0.75	0.708	0.703	0.0508	0.1123	0.703	0.0362	0.0790
0.2231	1.0	0.200	0.196	0.0299	0.0305	0.196	0.0213	0.0217
0.5108	1.0	0.400	0.398	0.0418	0.0423	0.396	0.0273	0.0300
0.9163	1.0	0.600	0.595	0.0458	0.0498	0.594	0.0310	0.0353
1.6094	1.0	0.800	0.794	0.0492	0.0537	0.793	0.0349	0.0382

Table 4.5: The averages and standard deviations (SD) of 500 estimates $\widehat{F}_{IPW}(t,u)$ and the averages of 500 standard errors (SE) at the selected points with the sample sizes (n) of 200 and 400, and the censoring rate of 50%

n			200			400		
t	u	$F(t,u)$	Mean	SD	SE	Mean	SD	SE
0.2231	0.25	0.158	0.157	0.0264	0.0274	0.158	0.0193	0.0193
0.5108	0.25	0.231	0.231	0.0325	0.0340	0.232	0.0229	0.0239
0.9163	0.25	0.247	0.247	0.0346	0.0371	0.248	0.0245	0.0260
1.6094	0.25	0.250	0.249	0.0350	0.0405	0.250	0.0249	0.0284
0.2231	0.5	0.196	0.195	0.0300	0.0300	0.195	0.0212	0.0211
0.5108	0.5	0.369	0.369	0.0400	0.0405	0.368	0.0273	0.0286
0.9163	0.5	0.469	0.469	0.0434	0.0481	0.470	0.0307	0.0340
1.6094	0.5	0.496	0.496	0.0458	0.0576	0.497	0.0325	0.0405
0.2231	0.75	0.200	0.199	0.0301	0.0302	0.199	0.0214	0.0213
0.5108	0.75	0.397	0.398	0.0419	0.0414	0.396	0.0273	0.0292
0.9163	0.75	0.581	0.581	0.0461	0.0503	0.581	0.0308	0.0355
1.6094	0.75	0.708	0.707	0.0512	0.0626	0.707	0.0365	0.0443
0.2231	1.0	0.200	0.200	0.0301	0.0302	0.199	0.0214	0.0213
0.5108	1.0	0.400	0.400	0.0419	0.0415	0.399	0.0274	0.0293
0.9163	1.0	0.600	0.600	0.0461	0.0503	0.600	0.0313	0.0356
1.6094	1.0	0.800	0.800	0.0494	0.0598	0.800	0.0349	0.0425

Table 4.6: The averages and standard deviations (SD) of 500 estimates $\widehat{F}_{IM}(t,u)$ at the selected points with the sample sizes (n) of 200 and 400, and the censoring rate of 50%

n			200		400	
t	u	$F(t,u)$	Mean	SD	Mean	SD
0.2231	0.25	0.158	0.157	0.0264	0.158	0.0193
0.5108	0.25	0.231	0.231	0.0324	0.231	0.0229
0.9163	0.25	0.247	0.245	0.0344	0.247	0.0244
1.6094	0.25	0.250	0.244	0.0345	0.247	0.0247
0.2231	0.5	0.196	0.195	0.0300	0.195	0.0211
0.5108	0.5	0.369	0.368	0.0399	0.368	0.0273
0.9163	0.5	0.469	0.467	0.0432	0.469	0.0306
1.6094	0.5	0.496	0.486	0.0450	0.492	0.0323
0.2231	0.75	0.200	0.199	0.0300	0.199	0.0213
0.5108	0.75	0.397	0.397	0.0418	0.396	0.0273
0.9163	0.75	0.581	0.578	0.0459	0.580	0.0308
1.6094	0.75	0.708	0.694	0.0507	0.701	0.0365
0.2231	1.0	0.200	0.199	0.0301	0.199	0.0214
0.5108	1.0	0.400	0.399	0.0419	0.398	0.0274
0.9163	1.0	0.600	0.597	0.0459	0.598	0.0312
1.6094	1.0	0.800	0.786	0.0490	0.793	0.0349

Table 4.7: The empirical coverage probabilities of $\widehat{F}_{HL}(t,u)$ at the selected points with the sample sizes (n) of 200 and 400, and the censoring rates (c.r.) of 30% and 50%

c.r.		30%		50%	
n		200	400	200	400
t	u				
0.2231	0.25	0.952	0.948	0.956	0.934
0.5108	0.25	0.958	0.978	0.964	0.948
0.9163	0.25	0.970	0.984	0.984	0.990
1.6094	0.25	0.990	1.000	0.998	1.000
0.2231	0.50	0.952	0.952	0.960	0.958
0.5108	0.50	0.976	0.966	0.972	0.966
0.9163	0.50	0.986	0.990	0.990	0.994
1.6094	0.50	1.000	1.000	1.000	1.000
0.2231	0.75	0.952	0.954	0.950	0.960
0.5108	0.75	0.968	0.972	0.954	0.958
0.9163	0.75	0.986	0.988	0.974	0.974
1.6094	0.75	1.000	1.000	0.998	1.000
0.2231	1.0	0.952	0.954	0.950	0.960
0.5108	1.0	0.968	0.970	0.958	0.952
0.9163	1.0	0.982	0.984	0.972	0.962
1.6094	1.0	0.994	0.976	0.970	0.968

Table 4.8: The empirical coverage probabilities of $\widehat{F}_{IPW}(t, u)$ at the selected points with the sample sizes (n) of 200 and 400, and the censoring rates (c.r.) of 30% and 50%

c.r.		30%		50%	
n		200	400	200	400
t	u				
0.2231	0.25	0.946	0.952	0.954	0.946
0.5108	0.25	0.968	0.934	0.952	0.942
0.9163	0.25	0.976	0.934	0.966	0.950
1.6094	0.25	0.978	0.944	0.974	0.976
0.2231	0.50	0.962	0.948	0.950	0.964
0.5108	0.50	0.960	0.950	0.970	0.966
0.9163	0.50	0.960	0.942	0.962	0.964
1.6094	0.50	0.970	0.950	0.988	0.982
0.2231	0.75	0.954	0.948	0.944	0.956
0.5108	0.75	0.966	0.940	0.958	0.962
0.9163	0.75	0.956	0.954	0.956	0.966
1.6094	0.75	0.970	0.964	0.976	0.984
0.2231	1.0	0.954	0.948	0.946	0.958
0.5108	1.0	0.970	0.936	0.956	0.966
0.9163	1.0	0.968	0.956	0.968	0.962
1.6094	1.0	0.964	0.964	0.968	0.978

Table 4.9: The averages and standard deviations (SD) of 500 estimates $\widehat{F}^*_{IPW}(t,u)$ and the averages of 500 standard errors (SE) at the selected points with the sample sizes (n) of 200 and 400, and the mixture rate of 30%

n			200			400		
t	u	$F^*(t,u)$	Mean	SD	SE	Mean	SD	SE
0.2231	0.25	0.159	0.158	0.0282	0.0288	0.160	0.0196	0.0204
0.5108	0.25	0.233	0.232	0.0326	0.0351	0.235	0.0239	0.0248
0.9163	0.25	0.249	0.247	0.0337	0.0364	0.250	0.0246	0.0257
1.6094	0.25	0.252	0.250	0.0340	0.0366	0.252	0.0248	0.0259
0.2231	0.50	0.198	0.197	0.0316	0.0321	0.198	0.0219	0.0226
0.5108	0.50	0.372	0.370	0.0418	0.0439	0.373	0.0311	0.0309
0.9163	0.50	0.472	0.470	0.0474	0.0488	0.473	0.0336	0.0344
1.6094	0.50	0.498	0.497	0.0498	0.0499	0.499	0.0341	0.0351
0.2231	0.75	0.202	0.200	0.0318	0.0324	0.201	0.0220	0.0229
0.5108	0.75	0.400	0.398	0.0424	0.0453	0.401	0.0327	0.0320
0.9163	0.75	0.584	0.582	0.0503	0.0528	0.586	0.0367	0.0372
1.6094	0.75	0.711	0.708	0.0538	0.0545	0.713	0.0385	0.0384
0.2231	1.0	0.202	0.201	0.0318	0.0324	0.202	0.0219	0.0229
0.5108	1.0	0.403	0.400	0.0425	0.0455	0.404	0.0328	0.0321
0.9163	1.0	0.603	0.600	0.0504	0.0533	0.604	0.0373	0.0375
1.6094	1.0	0.803	0.799	0.0530	0.0539	0.805	0.0385	0.0380

Table 4.10: The averages and the standard deviations (SD) of 500 estimates $\widehat{F}^*_{IPW}(t,u)$ and the averages of 500 standard errors (SE) at the selected points with the sample sizes (n) of 200 and 400, and the mixture rate of 50%

n			200			400		
t	u	$F^*(t,u)$	Mean	SD	SE	Mean	SD	SE
0.2231	0.25	0.204	0.205	0.0348	0.0374	0.203	0.0256	0.0265
0.5108	0.25	0.286	0.287	0.0420	0.0442	0.285	0.0299	0.0314
0.9163	0.25	0.301	0.301	0.0439	0.0453	0.299	0.0308	0.0323
1.6094	0.25	0.303	0.303	0.0443	0.0454	0.301	0.0311	0.0324
0.2231	0.50	0.252	0.254	0.0385	0.0414	0.250	0.0285	0.0293
0.5108	0.50	0.444	0.448	0.0514	0.0534	0.441	0.0374	0.0382
0.9163	0.50	0.535	0.538	0.0567	0.0569	0.533	0.0404	0.0411
1.6094	0.50	0.559	0.561	0.0570	0.0575	0.558	0.0396	0.0416
0.2231	0.75	0.257	0.259	0.0387	0.0418	0.256	0.0289	0.0296
0.5108	0.75	0.478	0.481	0.0521	0.0549	0.476	0.0383	0.0394
0.9163	0.75	0.662	0.664	0.0592	0.0596	0.661	0.0417	0.0435
1.6094	0.75	0.782	0.783	0.0605	0.0575	0.779	0.0425	0.0428
0.2231	1.0	0.257	0.259	0.0388	0.0418	0.256	0.0289	0.0297
0.5108	1.0	0.482	0.485	0.0523	0.0550	0.479	0.0384	0.0395
0.9163	1.0	0.683	0.686	0.0592	0.0598	0.682	0.0417	0.0437
1.6094	1.0	0.870	0.873	0.0573	0.0526	0.869	0.0405	0.0404

Table 4.11: The empirical coverage probabilities of $\widehat{F}^*_{IPW}(t,u)$ at the selected points with two sample sizes (n) of 200 and 400, and the mixture rates of censoring and death (m.r.) of 30% and 50%

m.r.		30%		50%	
n		200	400	200	400
t	u				
0.2231	0.25	0.948	0.950	0.964	0.948
0.5108	0.25	0.960	0.956	0.966	0.960
0.9163	0.25	0.970	0.958	0.954	0.958
1.6094	0.25	0.962	0.954	0.956	0.958
0.2231	0.50	0.950	0.960	0.956	0.950
0.5108	0.50	0.962	0.944	0.944	0.962
0.9163	0.50	0.958	0.954	0.946	0.952
1.6094	0.50	0.958	0.964	0.936	0.962
0.2231	0.75	0.944	0.956	0.956	0.958
0.5108	0.75	0.966	0.938	0.958	0.958
0.9163	0.75	0.956	0.944	0.944	0.956
1.6094	0.75	0.948	0.942	0.904	0.960
0.2231	1.0	0.944	0.960	0.956	0.958
0.5108	1.0	0.962	0.936	0.954	0.954
0.9163	1.0	0.952	0.946	0.942	0.958
1.6094	1.0	0.944	0.940	0.864	0.944

Chapter 5

Application to Colorectal Cancer Data

The used colorectal cancer data arise from the SEER-Medicare database. A total of 71,519 patients with the SEER registries were systematically recruited since January 1, 1983. The repeated medicare reimbursements (dollars) and the corresponding times (months) were recorded between January 1, 1983 and August 31, 1993. Here, we apply our proposed methods to estimate the distribution of first pair of medicare reimbursement and claiming time of patients. The baseline covariates age and cancer stage are considered in our analysis. Moreover, the time to colorectal cancer-related death and last follow-up are included. The stage variable is the American Joint Committee on Cancer (AJC) clinical stage of disease, which ranges from 0 to 4 according to the severity of disease. The age variable is further categorizd into three layers (61-70, 71-80, and >80). More detailed explorations of data can be found in Bang (2005).

In this chapter, a random sample of size about 2000 is selected and analyzed. The range of patients' age in this sample is mainly from 65 to 103 years old. The features of sub-sample table 5.1 show the representative of whole data. The aim of our study is to estimate the joint distributions of claiming time and medicare reimbursement under different age layers and clinical stages of disease. Moreover, the mean medicare reimbursement and the probabilities of claiming time occurring before the death time are evaluated. Evidenced by the numerical studies, the low mixture rate of censoring and death ($< 2\%$) in this sample will ensure the accuracy and precision of estimated distributions and related quantities. The results summarized in table 5.2 indicate that patients with older age or more severe disease stage tends to receive larger reimbursements from medicare. It is further detected that the greatest costs occur in the age layer of 71-80 and the disease stage 3. Those patients with older age or more severe disease stage are prone to claim reimbursements prior to death. The reason might be that older or less healthy patients tend to be ailing and raise medical expenditure. In table 5.3, the probabilities of claiming time prior to death are generally high, especially in the groups of older age and more severe disease stage. Patients with disease stage 4 receive the greatest medicare reimbursements in the age layer of 61-70, while the greatest reimbursements for patients with age more than 70 occur in the disease stage 3.

The patterns of joint distributions in various age layers and disease stages are displayed in figures 5.1-5.2. The marginal distribution of claiming time and reimbursement for patients with the first reimbursement prior to death are also presented in figures 5.3-5.4. Figure 5.3 reveals that the claiming times of patients with age

Table 5.1: The characteristics of the colorectal cancer data and subsample

		Full data	Sub-sample
Male		51.5%	52.5%
Female		48.5%	47.5%
Age	61-70	23.0%	22.1%
	71-80	47.4%	46.7%
	> 80	29.6%	31.2%
Stage	0	6.7%	7.8%
	1	22.6%	22.5%
	2	31.0%	30.4%
	3	22.5%	23.7%
	4	17.1%	15.8%

more than eighty are shorter than those for younger patients. As for the reimbursements, patients with age more than seventy receive more reimbursements than younger patients. In figure 5.4, no apparent difference between the estimated curves of claiming time is detected for various disease stages. In contrast, patients in disease stage 0 reasonably incur less reimbursements than those in the more severe stages.

Table 5.2: The estimates of $P(D > X^o)$ and $E(Y^o)$ under different age layers and disease stages

		$\widehat{P}(D > X^o)$	$\widehat{E}(Y^o)$
Age	61-70	0.977	4914.37
	71-80	0.990	5632.27
	> 80	0.997	5403.41
Stage	0	0.985	2168.10
	1	0.981	4935.75
	2	0.990	5561.14
	3	0.991	6436.52
	4	1.000	5876.16

Table 5.3: The estimates of $P(D > X^o)$ and $E(Y^o)$ under different age-disease stage groups

Age	Stage	$P(D > X^o)$	$E(Y^o)$
61-70	0	1.00	1649.8
	1	0.96	4293.4
	2	0.99	5605.1
	3	0.97	5326.5
	4	1.00	5788.9
71-80	0	0.99	2521.0
	1	0.99	5035.7
	2	0.98	5882.6
	3	1.00	6740.7
	4	1.00	6025.1
> 80	0	0.96	1967.8
	1	1.00	5363.2
	2	1.00	5077.1
	3	1.00	6837.3
	4	1.00	5621.1

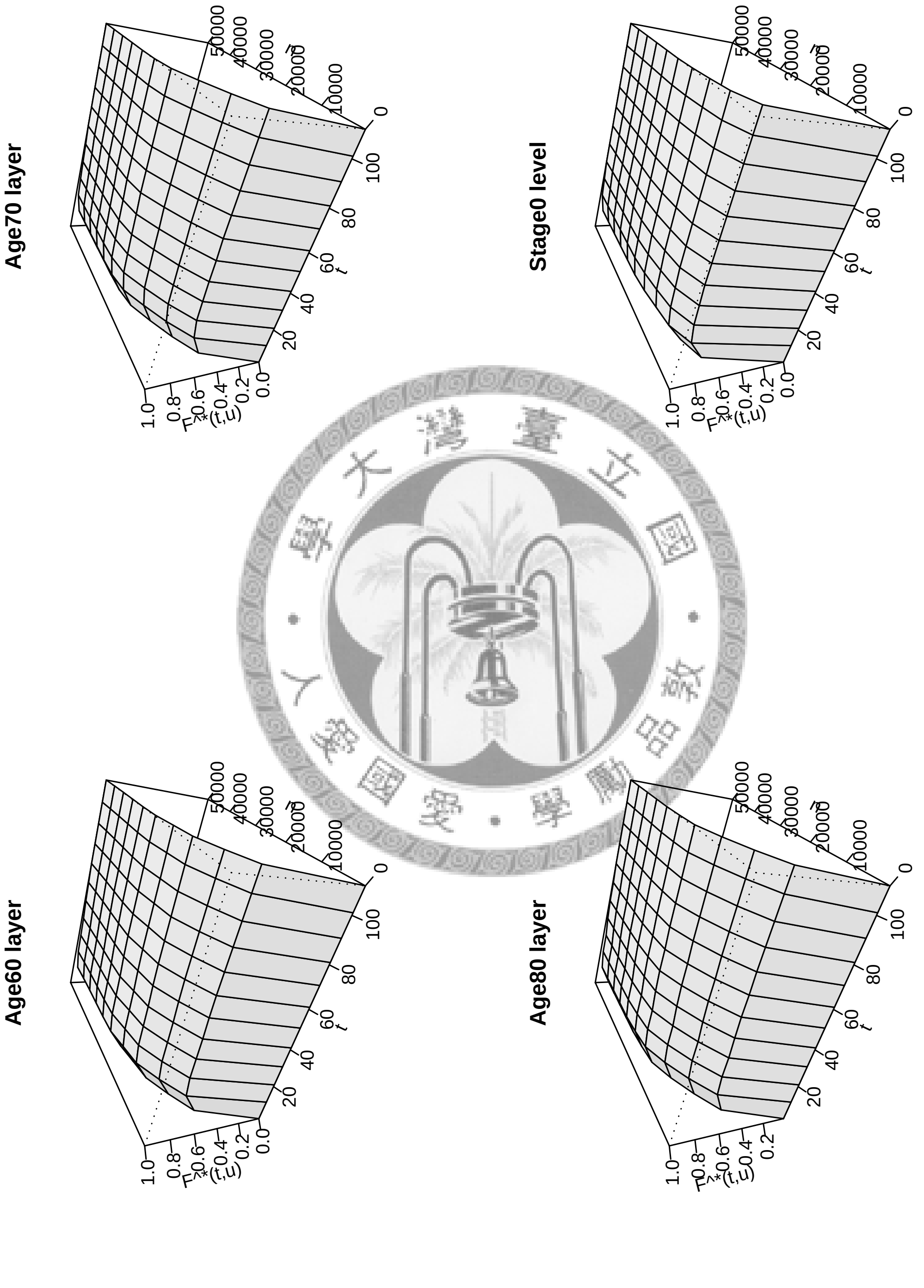

Figure 5.1: The joint distributions of claiming time and reimbursement for different age layers and disease stages

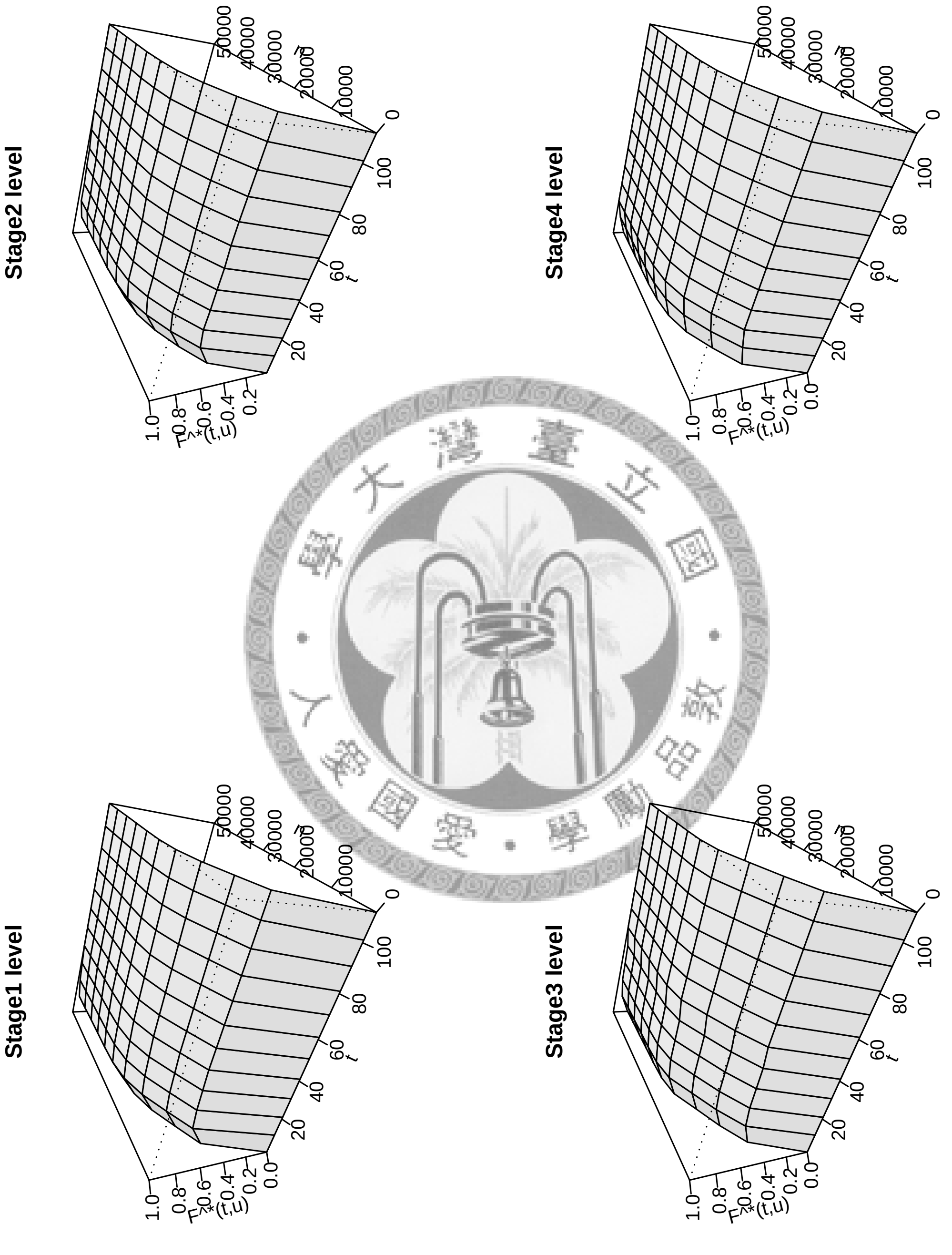

Figure 5.2: The joint distributions of claiming time and reimbursement for different age layers and disease stages

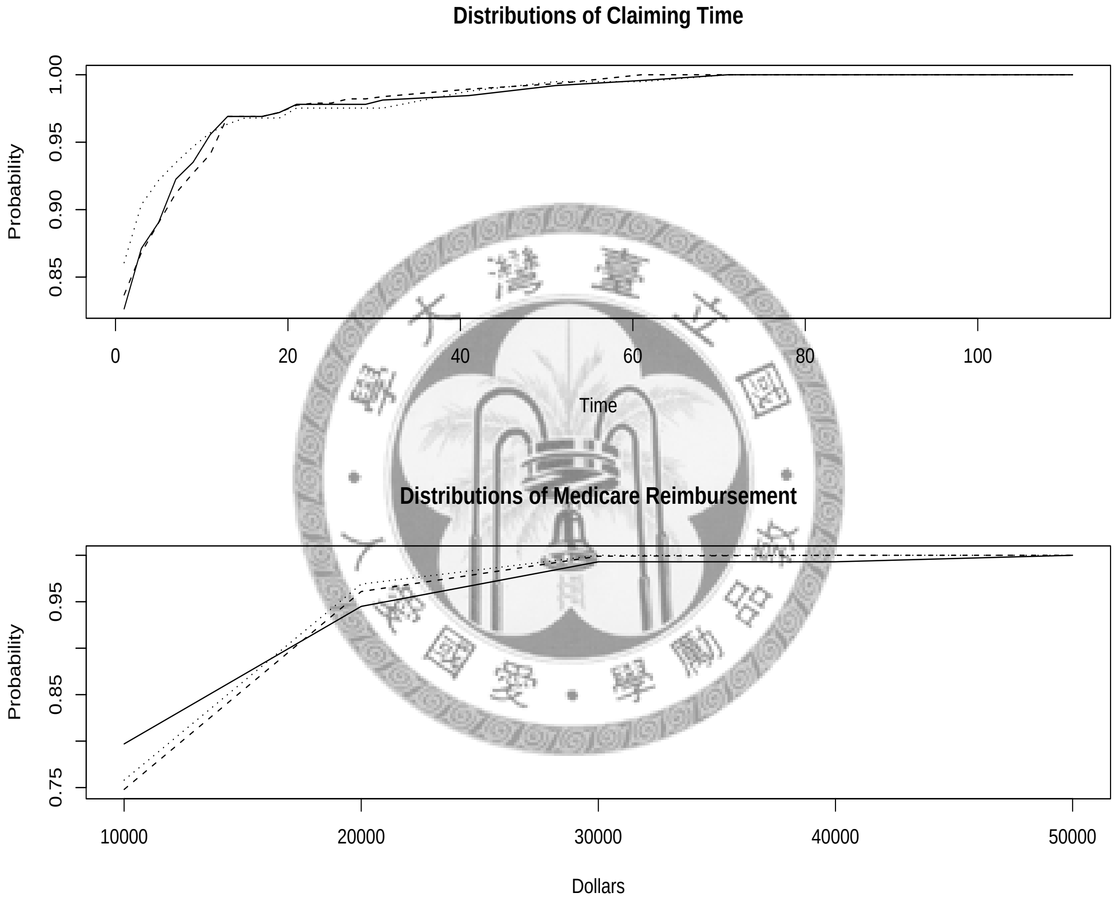

Figure 5.3: The estimates of $P(X^o \leq t|D > X^o)$ and $P(Y^o \leq u|D > X^o)$ for patients with age layers of 61-70(solid line), 71-80(dashed line) and > 80(dotted line)

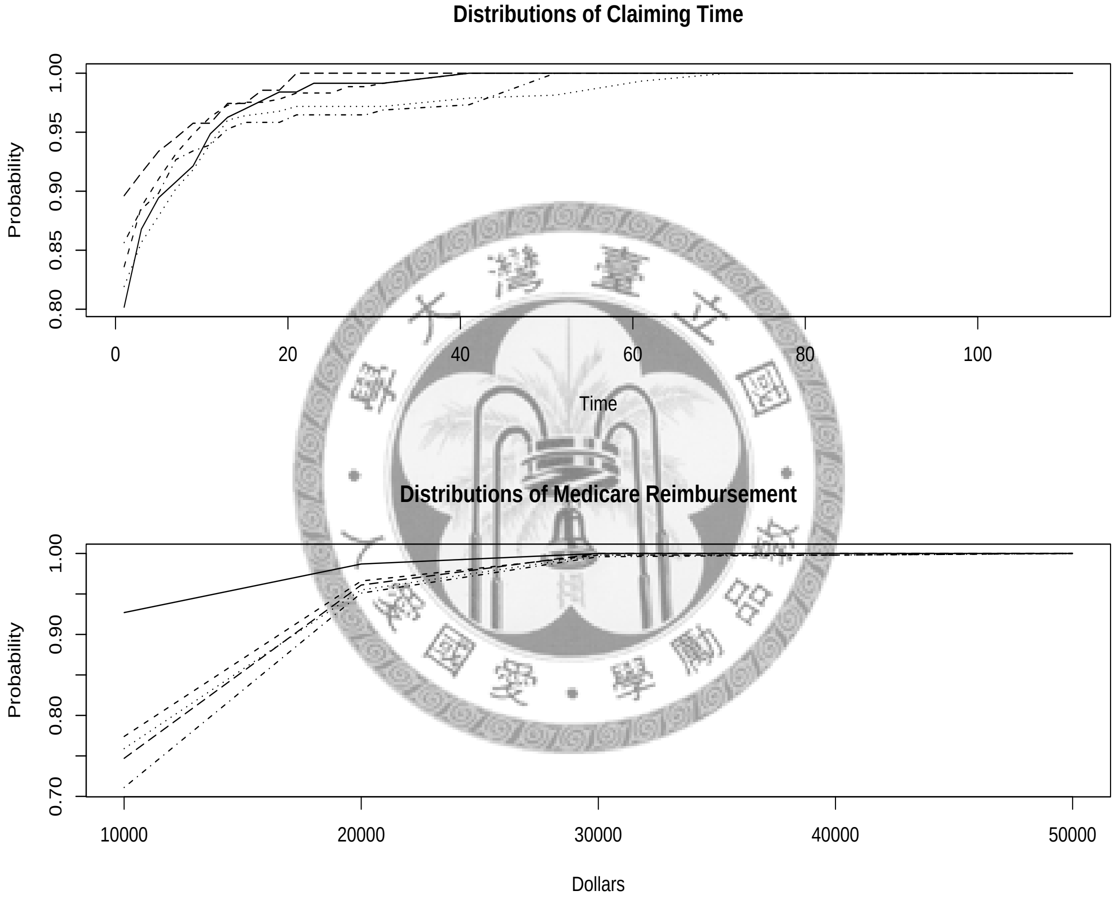

Figure 5.4: The estimates of $P(X^o \leq t|D > X^o)$ and $P(Y^o \leq u|D > X^o)$ for patients with disease stages of 0(solid line), 1(dashed line), 2(dotted line), 3(dotted-dash line) and 4(long-dashed line)

Chapter 6

Discussion

In this thesis, we propose several estimators for the joint distribution function of claiming time and medicare reimbursement based on two types of cost data. The limiting Gaussian processes of the estimators are also developed with the uniformly consistent estimators of the asymptotic variance-covariances. Without the occurrence of a terminal event, our numerical studies reveal that the IPW estimator surpasses the Huang-Louis and imputation estimators in computation cost. Moreover, the IPW estimation has more accurate estimator of the variance-covariance than the Huang Louis estimation. The performance of inference procedures are shown to be satisfactory.

In our application, an appropriate regression model would be helpful to investigate the influences of ages and disease stages on the joint distribution of claiming time and medicare reimbursement. The nonparametric IPW estimation approach will be reasonably extended to the estimation of parameters in the considered regression model. To solve the problem of asymmetric information or moral hazard in health economics, our further research will focus on seeking for the optimal composite factors to minimize the medical cost conditioning on the claiming time of interest. It is expected to help insurance companies to discriminate crafty policyholders.

As in the analyzed data, the claiming times and medicare reimbursements are intermittently occurring during the study period. Under the assumption that the recurrent pairs are independent and identically distributed conditioning on a latent variable, the estimation method of Huang and Wang (2005) can be applied to estimate the joint distribution of claiming time and medicare reimbursement. In biomedical contexts, this assumption seems to be out of reality. In our further study, we try to extend the developed methods to this issue with more suitable conditions being made.

Bibliography

[1] Akritas, M. G. and Van Keilegom, I. (2003). Estimation of bivariate and marginal distributions with censored data. *Journal of the Royal Statistical Society B* 65, 457-471.

[2] Akritas M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* 22, 1299-1327.

[3] Bang, H. (2005). Medical cost analysis: Application to colorectal cancer data from the SEER Medicare database. *Contemporary Clinical Trials* 26, 586-597.

[4] Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* 66, 429-436.

[5] Burke, M. D. (1988). Estimation of a bivariate distribution function under random censorship. *Biometrika* 75, 379-382.

[6] Campbell, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika* 68, 417-422.

[7] Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Annals of Statistics* 16, 1475-1489.

[8] Fleming, T. R. and Harrington, D. P. (1991). Counting process and survival analysis. New York: Wiley.

[9] Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika* 74, 549-555.

[10] Huang, C. Y. and Wang, M. C. (2005). Nonparametric estimation of the Bivariate Recurrence time distribution. *Biometrics* 61, 392-402.

[11] Huang, Y. J. and Louis. T. A.(1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika* 85, 785-798.

[12] Hudgens, M. G., Maathuis, M. H., and Gilbert, P. B. (2007). Nonparametric estimation of the joint distribution of a survival time subject to interval censoring and a continuous mark variable. *Biometrics* 63, 372-380.

[13] Lin, D. Y., Sun, W., and Ying, Z. L. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* 86, 59-70.

[14] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027-1057.

[15] Pollard, D. (1990). Empirical processes: Theory and applications. Hayward: Institute of Mathematical Statistics.

[16] Sherman, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics* 22, 439-459.

[17] Tsai, W. Y., Leurgans, S., and Crowley, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *Annals of Statistics* 14, 1351-1365.

[18] Wang, W. J. and Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika* 85, 561-572.