國立交通大學

統計研究所

碩士論文

半競爭風險資料受限於左截切的半母數推論

Semi-parametric Inference for Semi-competing Risks Data subject to Left Truncation

研究生：　林怡君

指導教授：王維菁　教授

中華民國九十六年六月

半競爭風險資料受限於左截切的半母數推論
Semi-parametric Inference for Semi-competing Risks Data subject to Left Truncation

研究生：林怡君	Student：Yichun Lin
指導教授：王維菁	Advisor：Weijing 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

June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

半競爭風險資料受限於左截切的半母數推論

研究生：林怡君　　　　　　　　　　指導教授：王維菁 教授

國立交通大學統計研究所碩士班

摘　要

本論文考慮以半母數推論方法估計在半競爭風險資料受限於左截切下的關連性。半競爭風險是多重事件發生的過程。以糖尿病的例子做說明，觀察值在罹患糖尿病之後往往會伴隨著一些併發症，如腎臟病或眼睛的病變而導致死亡，併發症與死亡之間的關係通常為醫學研究者所感興趣的。然而一些研究時間的限制，使得一些觀察值無法進入研究而被觀察到，我們稱這樣的觀察值受到截切。本論文的推論方法就是合併這兩個資料結構而進行，我們回顧了一些相關文獻，其中 Jiang 等人在 2005 年以 concordance 方法對同樣的資料結構做推論。而我們所提出的方法是以觀察值建立一系列 two-by-two 列聯表，此法可以視為 Clayton 在 1978 年對二維設限資料所提出的條件概似估計法的延伸。concordance 方法與 two-by-two 列聯表皆利用了 log-rank 的概念建立估計函數，我們將以模擬的結果比較 concordance 方法與我們所提出的方法。

關鍵字:半競爭風險資料；左截切；two-by-two 列聯表；設限。

Semi-parametric Inference for
Semi-competing Risks Data subject to Left Truncation

Student：Yichun Lin　　Advisor：Weijing Wang

Institute of Statistics
National Chiao Tung University

Abstract

The thesis considers semi-parametric inference for estimating the association parameter for a copula model based on semi-competing risks data which are further subject to left truncation. We review related literature including the paper by Jiang et al. (2005) who suggest solving the same problem by using concordant indicators. Alternatively we propose to construct an estimating function based on a series of two-by-two tables. Our method can be viewed as an extension of the conditional likelihood approach proposed by Clayton (1978) who originally considered bivariate censored data. Simulations are performed to assess the finite-sample performance.

Keywords: Archimedean copulas model; Left truncation; Semi-competing risks data; Two-by-two table.

謝　誌

本論文承蒙指導教授王維菁老師的指導，給我機會體驗到做研究的樂趣。老師的想法指引著本論文的進行，使我從中學習如何發揮創意於研究，尤其是論文架構的鋪陳及表達事情的能力更使我學到不少。而對於我的生活及未來規劃，老師也提供許多寶貴的意見，適時地給我指點迷津，幫助我解決問題，在此謹表由衷的感謝。另外要特別感謝的是博士班江村剛志學長，學長在研究過程中全心全力輔助老師教導著我，以循序漸進的方式帶領著我由基礎的觀念進入艱深的領域，對於問題盲點巨細靡遺不厭其煩地深入研究，這也是本論文能得以順利進行的重要因素。

在交大統計所的這兩年，除了在學業上獲益良多之外，班上同學也帶給了我以往不同的體驗。統計所 94 級是個很特別的一班，集結來自各校的精英，每位同學皆具有各自獨特的風格，一致的是大家樂觀的想法，學業上的積極，感謝他們陪伴著我讓我在這兩年成長不少。最後是我最摯愛的家人，爸媽從小到大一直鼓勵著我唸書，在我的求學過程遇上任何挫敗時，爸媽就是我的動力推使我樂觀向前，在寫下這篇誌謝的同時就意味著我的學業也告一段落，在此向他們說聲辛苦了，今後我會更加地努力不負他們的期待。

僅將誌謝獻給每一個曾經給我鼓勵的你們

林怡君　謹誌于
國立交通大學統計研究所
中華民國九十六年六月

Table of Contents

Chapter 1 Introduction

1.1 Background

In the thesis, we consider semi-parametric inference for Archimedean copulas models based on semi-competing risks data subject to left truncation. The Archimedean copulas (AC) family is a popular sub-class of copula models which have been frequently used to model bivariate failure-time variables. Copula models have the nice feature that the dependence structure can be studied separately from marginal analysis. Semi-parametric inference methods for estimating the association parameter without specifying the marginal distributions have been applied to different types of incomplete data. This research direction has brought substantial attentions due to its wide applicability and theoretical attractiveness.

Early work focused on bivariate censored data. The landmark paper by Clayton (1978) proposed a useful copula model and a semi-parametric inference procedure for estimating the association parameter. Specifically Clayton’s proposal is based on a conditional likelihood that measures the association on selected grid points without making any assumption on the marginal distributions. This approach was later shown to have a direct relationship with two-by-two tables constructed based on the grid points. Under the same model assumption, Oakes (1982, 1986) proposed to estimate the association parameter by utilizing the concordant information provided by paired observations. These two approaches have been further extended to more complicated data structures. For example, Fine et al. (2001) adapted Oakes’ (1986) closed-form estimator to semi-competing risks data. Jiang et al (2005) proposed the estimating function under semi-competing risks data subject to left truncation.

1.2 Overview of the thesis

In this thesis, we consider the same type of data structure as in Jiang et al. (2005) and propose a different approach by constructing an estimating function based a series of two-by-two tables, our idea can be viewed as an extension of the conditional likelihood proposed by Clayton (1978).

The outline of the thesis is summarized as follows. In Section 2.1, we introduce three types of data structure. The data type that we will study later is a combination of these three data types. We first introduce typical bivariate data, semi-competing risks data, and then truncation data. The related results developed for those three data types are discussed in Section 2.2. Chapter 3 contains a review of different inference methods for estimating the association parameter of a copula model. The conditional likelihood approach proposed by Clayton (1978), which was developed for bivariate right censored data, is discussed in Section 3.1. In Section 3.2, we study estimating functions constructed based on concordant indicators for bivariate right censored data (Oakes, 1986) and semi-competing risks data (Fine et al., 2001), respectively. In Section 3.3, we study the method using the information provided by a series of two-by-two tables (Day et al., 1997; Wang, 2003). Chapter 4 contains the main results of the thesis in which semi-competing risk data subject to left truncation is of interest. After introducing the concordance approach by Jiang et al. (2005) in Section 4.2, we present our proposal in Section 4.3. The modification of the proposed method for censored data will be discussed in Section 4.3. The results of simulation are showed in Chapter 5, which are divided two parts without external censoring and with external censoring, respectively. Chapter 6 contains some concluding remarks.

Chapter 2 Literature Review

Let (X,Y) be a pair of failure time variables which may be correlated. Sometimes due to the constraint of the observational scheme, these two variables may be subject to censoring or truncation. In Section 2.1, we introduce three different data structures which are commonly seen in applications. The data structure that we will study later is a combination of these three data structures. The related results of estimation for those three data structures are studied in Section 2.2.

2.1 Three Data Structures

To simplify the presentation, we may use the same notation for quantities with similar meanings under different data structures. For example, we use X' to denote the observed version of X with an indicator $\delta = I(X' = X)$. However the condition that $\delta = 1$ is different for different data structures.

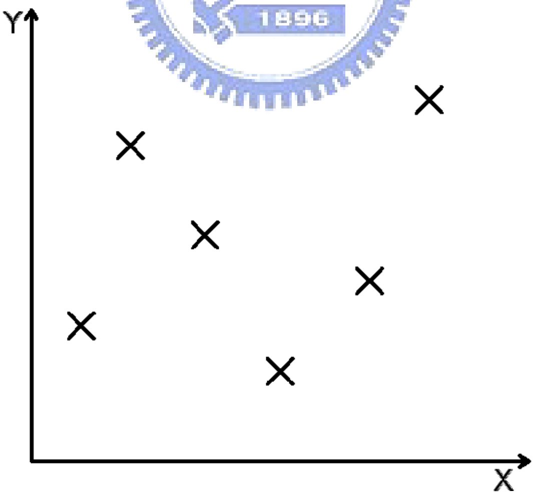


Figure 2.1: Typical Bivariate Data

A. Typical Bivariate Data

Suppose that (X,Y) represent lifetimes of twins or failure times occurred to paired organs. For the former, the dependence can be attributed to shared genetic or environmental factors. For the latter, the relationship may be explained by the same internal biological system of the subject. Figure 2.1 depicts such data in which replications of (X,Y) have no specific restriction.

External censoring may occur to each member of the pair. Let (C_1,C_2) be external censoring times so that one only observes $(X',Y',\delta_1,\delta_2)$ such that $X' = X \wedge C_1$, $Y' = Y \wedge C_2$, $\delta_1 = I(X < C_1)$ and $\delta_2 = I(Y < C_2)$, where $\wedge$ denotes the minimum and $I(\cdot)$ is the indicator function. It is usually assumed that (C_1,C_2) are independent of (X,Y). The observed data can be expressed as $\{(X_i',Y_i',\delta_{1i},\delta_{2i}),(i = 1,....,n)\}$, where $X_i' = X_i \wedge C_{1i}$, $Y_i' = Y_i \wedge C_{2i}$, $\delta_{1i} = I(X_i < C_{1i})$ and $\delta_{2i} = I(Y_i < C_{2i})$, are random replicates of $(X',Y',\delta_1,\delta_2)$. This type of data is most commonly seen in the literature of survival analysis.

B. Bivariate Analysis — Semi-competing Risks Data

Consider that (X,Y) represent the time to morbidity and the time to mortality of a specific disease on the same subject, respectively. Hence X is subject to right censoring by Y but not vice versa. Temporarily we ignore external censoring. Figure 2.2 depicts the structure of semi-competing risks data. Notice that observations of (X,Y) are located on the upper wedge. For those with $X > Y$, we only observe $(X \wedge \mathrm{Y} = Y, Y)$ which is located on the diagonal line. This type of data is called “semi-competing risks data” by Fine et al. (2001).

When external censoring occurs, it is reasonable to set $C_1 = C_2 = C$ since (X,Y) represent different failure times on the same subject. Usually it is assumed that C is independent of (X,Y). When X is right censored by $Y \wedge C$ and Y is right censored by C, the observed data can be written as $\{(X_i',Y_i',\eta_i,\delta_i),(i = 1,....,n)\}$, where $X_i' = X_i \wedge Y_i \wedge C_i$, $Y_i' = Y_i \wedge C_i$, $\eta_i = I(X_i < (Y_i \wedge C_i))$, and $\delta_i = I(Y_i < C_i)$.

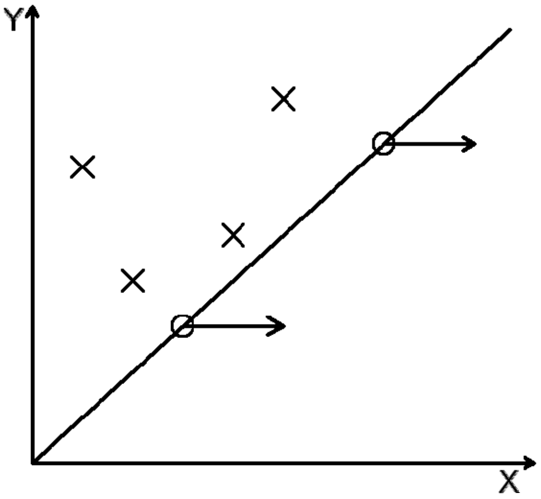


Figure 2.2: Semi-competing risks Data

C. Truncation Data

Here we consider a pair of failure times (Y, A) which have a truncation relationship. Specifically we can observe (Y, A) only if $Y > A$. We can say that Y is subject to left truncation by A while A is subject to right truncation by Y. Note that, unlike semi-competing risks data, we have no information when $Y < A$. In Figure 2.3, observations on the lower wedge will be completely missing and even their existence is unknown. Many applications consider left truncation in which Y is the variable of interest which is subject to truncation by A.

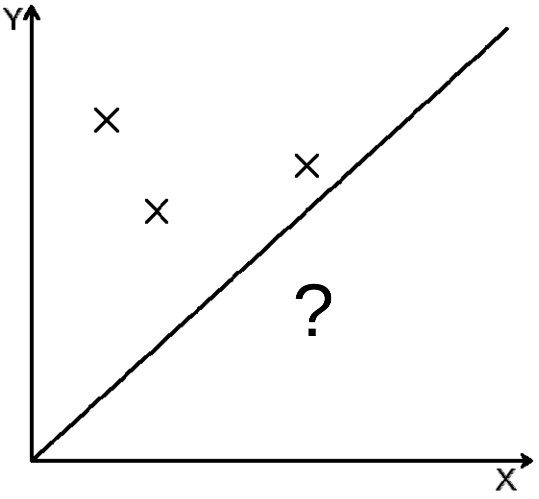


Figure 2.3: Truncation Data

2.2 Nonparametric Inferences under Three Data Structures

A. Typical Bivariate Data

For univariate survival data, Kaplan and Meier (1958) expressed the survival function as a product integral of the cumulative hazard function,

$$\Pr(T > t) = \prod_{u \le t} \left\{ 1 - \frac{\Pr(T \in [u, u + du))}{\Pr(T \ge u)} \right\}. \tag{2.1}$$

For the bivariate case, the Kaplan-Meier estimator can be still applied to estimate each of $\Pr(X > x)$ and $\Pr(Y > y)$, respectively. When censoring occurs, $\Pr(X > x)$ and $\Pr(Y > y)$ are estimated based on $\{(X_i', \delta_{1i}), (i = 1, \ldots, n)\}$ and $\{(Y_i', \delta_{2i}), (i = 1, \ldots, n)\}$, respectively. The K-M estimator of $\Pr(X > t)$ and $\Pr(Y > t)$ are

$$\hat{\Pr}(X > t) = \prod_{u \le t} \left\{ 1 - \frac{\sum_{i=1}^{n} I(X_i' = t, \delta_{1i} = 1)}{\sum_{i=1}^{n} I(X_i' \ge t)} \right\},$$

$$\hat{\Pr}(Y > t) = \prod_{u \le t} \left\{ 1 - \frac{\sum_{i=1}^{n} I(Y_i' = t, \delta_{2i} = 1)}{\sum_{i=1}^{n} I(Y_i' \ge t)} \right\}.$$

Let $S(x, y)$ be the joint survival function of X and Y, where $S(x, y)$ $= \Pr(X > x, Y > y)$. There exist several nonparametric estimators of the joint survival function $S(x, y)$. The most well-known one was proposed by Dabrowska (1988).

B. Bivariate Analysis — Semi-competing risks data

With semi-competing risks data, the Kaplan-Meier estimator of $\Pr(Y > y)$ is still valid. However, due to dependent censoring, the Kaplan-Meier estimator of $\Pr(X > x)$ is biased. Actually the distribution of X is not identifiable nonparametrically.

Suppose that the external censoring is taken into account**,** one can estimate $S(x, y)$ for $x < y$ nonparametrically by

$$\hat{S}(x,y)=\frac{\sum_{i=1}^{n} I(X_i' \ge x, Y_i' \ge y)}{n\hat{G}(y)}, \tag{2.2}$$

where

$$\hat{G}(y)=\prod_{u\le y}\left\{1-\frac{\sum_{i=1}^{n} I(Y_i' = u, \delta_i = 0)}{\sum_{i=1}^{n} I(Y_i' \ge u)}\right\}.$$

However, to recover the dependence structure between X and Y, we not only need a valid estimator of $S(x,y)$ but also both of the marginal estimators as well. This implies that the dependence structure can not be recovered non-parametrically for semi-competing risks data. Therefore most authors have considered semi-parametric inference to investigate the dependence structure. The common model assumption is the Archimedean copula family which will be discussed in Section 2.3. We also adopt this approach in the thesis.

C. Truncation Data

Estimating the survival function of Y conditional on $Y > A$ by the Kaplan-Meier estimator may be biased. Under truncation, we observe $\{(Y_i, A_i), (i = 1,...,n)\}$ only if $Y_i > A_i$. Hence $\{Y_1,...,Y_n\}$ is no longer a random sample of Y.

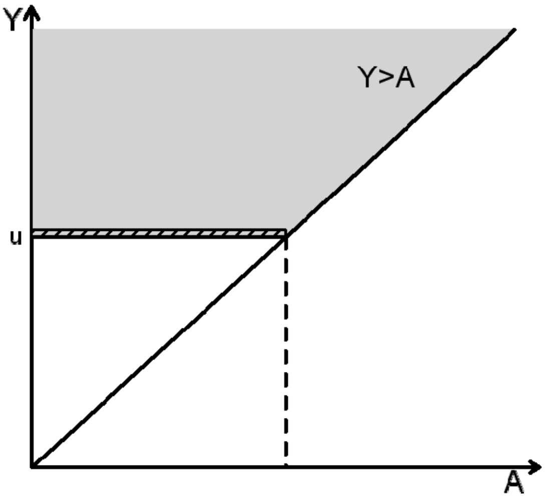


Figure 2.4 Risk Set for Truncation Data

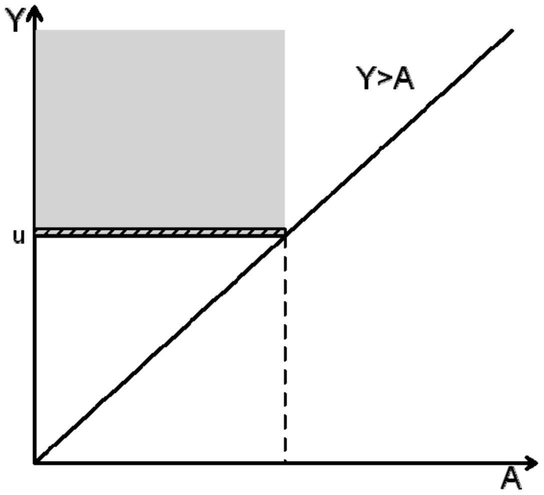


Figure 2.5: Modified Risk Set for Truncation Data

Figure 2.4 explains the truncation mechanism on the hazard estimation. After truncation, the original risk set $\{Y \geq u\}$ becomes $\{Y \geq u, Y > A\}$, the shaded area on Figure 2.4. The failure region $\{Y = u\}$ changes to $\{Y = u, Y > A\}$, the slash area on the figure.

If we use the set $\{Y \geq u, Y > A\}$ to be the new risk set and $\{Y = u, Y > A\}$ to be the new instantaneous risk set, it follows that

$$\frac{\sum_{i=1}^{n} I(Y_i = u)}{\sum_{i=1}^{n} I(Y_i \geq u)} \xrightarrow[p]{} \frac{\Pr(Y \in [u, u+du), Y > A)}{\Pr(Y \geq u, Y > A)} \neq \frac{\Pr(Y \in [u, u+du))}{\Pr(Y \geq u)}.$$

The resulting estimator of $\Pr(Y > t)$ tends to over-estimate the true survival function. To correct this bias, Lynden-Bell modified the set $\{Y \geq u, Y > A\}$ by cutting the set further. Lynden-Bell proposed that the new risk set is $\{Y \geq u, A < u\}$, the shaded area on Figure 2.5. The new instantaneous risk set is $\{Y = u, A < u\}$, the slash area on the figure.

Under the assumption that Y and A are independent, one can show that

$$\frac{\Pr(Y = u, u \geq A)}{\Pr(Y \geq u, u \geq A)} = \frac{\Pr(Y = u)}{\Pr(Y \geq u)}.$$

Consequently the modified estimator proposed by Lynden-Bell,

$$\hat{\Pr}(Y>t)=\prod_{u\le t}\left\{1-\frac{\sum_{i=1}^{n} I(Y_i=u,u>A_i)}{\sum_{i=1}^{n} I(Y_i\ge u,u>A_i)}\right\}, \qquad (2.3)$$

is a valid estimator for $\Pr(Y>t)$.

Suppose that there is an external censoring variable C for Y. It is assumed that C is independent of Y. The observed data are $\{Y_i', A_i,\delta_i),(i=1,......,n)\}$ conditional on $Y_i'>A_i$, where $Y_i'=Y_i\wedge C_i$ and $\delta_i=I(Y_i<C_i)$. It follows that

$$\frac{\Pr(Y'=u,\delta=1,u>A)}{\Pr(Y'\ge u,u>A)}=\frac{\Pr(Y=u,u>A)}{\Pr(Y\ge u,u>A)}=\frac{\Pr(Y=u)}{\Pr(Y\ge u)}.$$

The resulting Lynden-Bell's estimator becomes

$$\hat{\Pr}(Y>t)=\prod_{u\le t}\left\{1-\frac{\sum_{i=1}^{n} I(Y_i'=u,\delta_i=1,u>A_i)}{\sum_{i=1}^{n} I(Y_i'\ge u,u>A_i)}\right\}. \qquad (2.4)$$

Tsai (1991) claimed that most existing procedures for truncation data are still correct under a weaker assumption of quasi-independence, and then proposed a test to verify this condition. The recent paper by Chaieb et al. (2006) consider the assumption that Y and A are correlated and proposed a semi-parametric inference procedure under a "semi-survival" Archimedean copula model. Note that in the thesis, we only assume quasi-independence between the truncation time A and the survival time Y.

2.3 Copula Models and Archimedean Copula Model

Copula models are often used to describe the association between two failure time variables. For the bivariate case, a copula function can be written as $C(u,v)$, which may be parameterized as $C_\alpha(u,v)$ for $u,v\in[0,1]$. The Archimedean copula (AC) family is a subclass of copula models. A copula is said to be Archimedean copula (AC) if it can be expressed in the following form,

$$C_\alpha(u,v)=\phi_\alpha^{-1}\{\phi_\alpha(u)+\phi_\alpha(v)\}, \qquad (2.5)$$

where $\phi_\alpha : [0,1] \to [0,\infty]$ satisfying $\phi_\alpha(1) = 0$, $\varphi'_\alpha(t) < 0$ and $\phi''_\alpha(t) > 0$. Note that the AC family simplifies the bivariate relationship via the univariate function $\phi_\alpha(\cdot)$. The function $\phi_\alpha(\cdot)$ is the generator of the copula. Important proporties of AC models have been derived in Genest et al. (1986), Oakes (1989) and Genest et al. (1993).

One of the most well-known AC model is the Clayton model with $\phi_\alpha(t) = (t^{-\alpha} - 1)/\alpha$ for some $\alpha > 0$, then

$$C_\alpha(u,v) = \{u^{-\alpha} + v^{-\alpha} - 1\}^{-1/\alpha}. \tag{2.6}$$

In applications, the copula structure is imposed on (X,Y) such that one can write

$$S(x,y) = C_\alpha\{\Pr(X > x), \Pr(Y > y)\} \tag{2.7}$$

Accordingly an AC model defined on the joint survival function can be written as

$$S(x,y) = \phi_\alpha^{-1}[\phi_\alpha\{\Pr(X > x)\} + \phi_\alpha\{\Pr(Y > y)\}].$$

The AC family has nice analytic properties which are useful for further statistical inference. For example, consider the odds ratio function proposed by Oakes (1989):

$$\theta^*(x,y) = \frac{S(x,y) \cdot \partial^2 S(x,y) / \partial x \partial y}{\partial S(x,y) / \partial x \cdot \partial S(x,y) / \partial y} \tag{2.8}$$

For an AC model $\theta^*(x,y)$ can be simplified as $\theta_\alpha(S(x,y))$, where

$$\theta^*_\alpha(v) = -v\phi''(v) / \phi'_\alpha(v). \tag{2.9}$$

Chapter 3 Review of Semi-Parametric Analysis

In this chapter, we review three semi-parametric inference methods for estimating the copula association parameter under different data structures. In Section 3.1 and 3.2, each inference method will be first applied to bivariate data without censoring and then to a more general situation including external censoring. In addition, the application to semi-competing risks data is also considered. In Section 3.3, two-by-two table method by Wang (2003) is directly applied to semi-competing risks data without censoring. Without losing generality, we assume there is no tie in the sample for the observed data.

3.1 The Conditional Likelihood Approach

This approach was first proposed by Clayton in his landmark paper (Clayton, 1978). To simplify the discussion and without loss of generality, we temporarily assume that a random sample of (X,Y) can be observed without censoring and denoted as $\{(X_i,Y_i),(i=1,..,n)\}$. Clayton (1978) defined the following set of grid points, denoted as φ, such that

$$\varphi=\left\{(x,y):\sum_{i=1}^{n}I(X_i=x,Y_i\geq y)=1,\sum_{i=1}^{n}I(X_i\geq x,Y_i=y)=1\,|\,0<x,y\leq\infty\right\}. \tag{3.1}$$

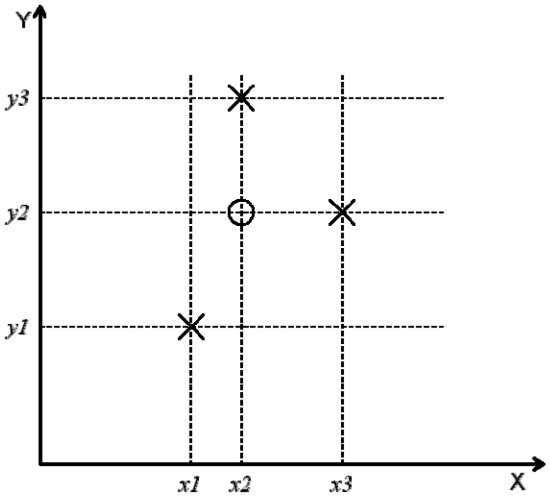


Figure 3.1: An Example of The Set φ for An Artificial Data Set

Figure 3.1 depicts the set φ for an artificial data set which consists of three observations,

(x_1, y_1), (x_2, y_3), and (x_3, y_2), marked as “×”. According to the definition, the set φ consists of four grid points (x_1, y_1), (x_2, y_3), (x_3, y_2), and (x_2, y_2). Note that (x_2, y_2) on Figure 3.1 is not an observed failure point. We mark such a point by “○”. Define

$$\Delta(x,y) = \begin{cases} 1, \text{ if } \sum_{i=1}^{n} I(X_i = x, Y_i = y) = 1; \\ 0, \text{ if } \sum_{i=1}^{n} I(X_i = x, Y_i > y) = 1 \text{ and } \sum_{i=1}^{n} I(X_i > x, Y_i = y) = 1. \end{cases} \qquad 3.2)$$

Note that $\Delta(\mathrm{x,y}) = 1$ implies the grid point (x,y) is associated with an observation of $(X,\ Y)$ (i.e. a point marked by “×”). If $\Delta(x,y) = 0$, the point (x,y) is not an observed point (i.e. (x_2, y_2) in the above example and marked by “○”). Define $R(x,y) = \sum_{i=1}^{n} I(X_i \ge x, Y_i \ge y)$ which counts the number at risk at time (x,y). Conditional on the value of $R(x,y)$, $\Delta(x,y)$ follows a Bernoulli distribution with the probability $\Pr\{\Delta(x,y) = 1 \mid (x,y) \in \varphi, R(x,y)\}$. For an AC model,

$$\Pr\{\Delta(x,y) = 1 \mid (x,y) \in \varphi, R(x,y)\} = \frac{\theta_\alpha(S(x,y))}{\theta_\alpha(S(x,y)) + R(x,y) - 1}. \qquad (3.3)$$

For the Clayton model with $\theta_\alpha(S(x,y)) = \alpha$, we have

$$\Pr\{\Delta(x,y) = 1 \mid (x,y) \in \varphi, R(x,y)\} = \frac{\alpha}{\alpha + R(x,y) - 1}, \qquad (3.4)$$

which does not involve the nuisance parameter $S(x,y)$.

Clayton (1978) suggested that the distribution of $\Pr\{R(x,y) = r \mid (x,y) \in \varphi\}$ may be ignored in the likelihood construction since it may contain only little information about α. Under a working assumption that $\Delta(x,y)$ and $\Delta(x',y')$ are independent for different grid points (x,y) and $(x',y') \in \varphi$, the “conditional” likelihood for an AC model can be written as the product over the conditional probabilities of $\Delta(x,y)$ for all grid points in the set φ. Specifically

$$L(\alpha, S(x,y)) = \prod_{(x,y)\in\varphi} \left[\Pr\{\Delta(x,y) = 1 \mid (x,y) \in \varphi, R(x,y)\}\right]^{\Delta(x,y)} \times \left[\Pr\{\Delta(x,y) = 0 \mid (x,y) \in \varphi, R(x,y)\}\right]^{1-\Delta(x,y)}. \qquad (3.5)$$

For the Clayton model, the corresponding log-likelihood is given by

$$\ell(\alpha)=\sum_{(x,y)\in\varphi}\left[\Delta(x,y)\log\left(\frac{\alpha}{R(x,y)-1+\alpha}\right)+\{1-\Delta(x,y)\}\log\left(\frac{R(x,y)-1}{R(x,y)-1+\alpha}\right)\right]. \tag{3.6}$$

So that the estimating equation becomes

$$\frac{\partial\ell(\eta)}{\partial\eta}=\sum_{(x,y)\in\varphi}\left\{\Delta(x,y)-\frac{\alpha}{R(x,y)-1+\alpha}\right\}, \tag{3.7}$$

where $\eta=\log\alpha$. The solution can be denoted as

$$\hat{\alpha}_L=\frac{\sum_{(x,y)\in\varphi}\{R(x,y)-1\}\times\Delta(x,y)}{\sum_{(x,y)\in\varphi}\{1-\Delta(x,y)\}}.$$

For general AC models, we can maximize $L(\alpha,\hat{S}(x,y))$, where $\hat{S}(x,y)$ is the empirical estimator of $S(x,y)$. However the resulting estimator of α may not have an explicit formula.

When censoring is taken into account, the set φ can be modified as

$$\left\{(x,y):\sum_{i=1}^{n}I(X_i'=x,Y_i'\geq y,\delta_{1i}=1)=1,\sum_{i=1}^{n}I(X_i'\geq x,Y_i'=y,\delta_{2i}=1)=1\right\}. \tag{3.8}$$

The definition of $\Delta(x,y)$ is changed to

$$\Delta(x,y)=\begin{cases}1,\ \text{if}\ \sum_{i=1}^{n}I(X_i'=x,Y_i'=y,\delta_{1i}=\delta_{2i}=1)=1;\\ 0,\ \text{if}\ \sum_{i=1}^{n}I(X_i'=x,Y_i'>y,\delta_{1i}=1)=1,\sum_{i=1}^{n}I(X_i'>x,Y_i'=y,\delta_{2i}=1)=1.\end{cases} \tag{3.9}$$

The resulting estimating function under AC model involves the plugged-in estimator, $\hat{S}(x,y)$, which can be the Dabrowska's estimator. The estimator can be modified accordingly. The same principle can also be applied to different data structures, such as semi-competing risks data, in which

$$\hat{S}(x,y)=\frac{\sum_{i=1}^{n}I(X_i'\geq x,Y_i'\geq y)}{n\hat{G}(y)}.$$

3.2 Estimating Functions Based on Concordance Indicators

Let (X_i, Y_i) and (X_j, Y_j) be independent replications of (X, Y). Define the indicator, $\Delta_{ij} = I\{(X_i\text{-}X_j)(Y_i\text{-}Y_j) > 0\}$. The two pairs are said to be concordant if Δ_{ij}=1 and discordant if Δ_{ij}=0. This indicator reveals dependence relationship between X and Y. Oakes (1989) proposed the following time-dependent association measure：

$$\theta(x,y) = \frac{\Pr(\Delta_{ij} = 1 \mid \tilde{X}_{ij} = x, \tilde{Y}_{ij} = y)}{\Pr(\Delta_{ij} = 0 \mid \tilde{X}_{ij} = x, \tilde{Y}_{ij} = y)}, \qquad 0 < x, y \le \infty \tag{3.10}$$

where $\tilde{X}_{ij} = X_i \wedge X_j$ and $\tilde{Y}_{ij} = Y_i \wedge Y_j$. For Clayton's model, we can find that $\theta(x,y) = \alpha$ for $0 < x, y \le \infty$ and $E(\Delta_{ij} \mid \tilde{X}_{ij} = x, \tilde{Y}_{ij} = y) = \alpha/(\alpha+1)$. The information can be utilized in the inference of α. Assuming the Clayton model, Oakes (1982) proposed the estimating function,

$$U(\alpha) = \sum_{i=1}^{n} \sum_{j;j>i} \left(\Delta_{ij} - \frac{\alpha}{\alpha+1} \right). \tag{3.11}$$

The solution can be written as

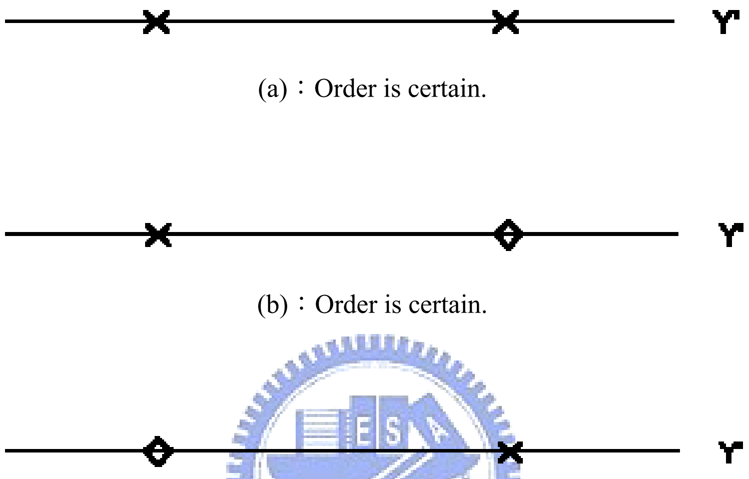
$$\hat{\alpha}_C = \frac{\sum_{i=1}^{n} \sum_{j;j>i} \Delta_{ij}}{\sum_{i=1}^{n} \sum_{j;j>i} (1 - \Delta_{ij})}$$

Note that the concordant estimator $\hat{\alpha}_C$ is a U-statistic which is useful in the establishment of large-sample theory.

To further extend the above idea to incomplete data, the challenge is that some values of Δ_{ij} may be uncertain due to censoring. The following discussion is about how the effects of censoring affects the information of Δ_{ij}. We can write

$$\Delta_{ij} = I\{(X_i - X_j) > 0\} \times I\{(Y_i - Y_j) > 0\} + I\{(X_i - X_j) < 0\} \times I\{(Y_i - Y_j) < 0\}. \tag{3.12}$$

This means that to know the value of Δ_{ij}, we need to know the marginal orders of both (X_i, X_j) and (Y_i, Y_j). Given (Y_i', δ_{2i}) and (Y_j', δ_{2j}), the order of (Y_i, Y_j) is certain if $Y_i' \wedge Y_j'$ is associated with an uncensored observation. The phenomenon can be explained by the following figures：



(a)：Order is certain.

(b)：Order is certain.

(c)：Order is uncertain.

× ：$\delta = 1$， ◇ ：$\delta = 0$

Figure 3.2 : The Effect of Censoring on the Order of Two Pairs

Notice that in Figure 3.2 (a) and Figure 3.2 (b), the orders of the two pairs are certain, while in Figure 3.2 (c) is not. Define that $\tilde{Y}_{ij}' = Y_i' \wedge Y_j'$ which is observed if only if it is smaller than both $C_{2.i}$ and C_{2j}. Mathematically, we can write the above condition as $\tilde{Y}_{ij}' < \tilde{C}_{2ij}$, where $\tilde{C}_{2ij} = C_{2i} \wedge C_{2j}$. Similar conclusions can be applied to determine the order of X_i and X_j. As long as $\tilde{X}_{ij}' < \tilde{C}_{1ij}$, where $\tilde{X}_{ij}' = X_i' \wedge X_j'$ and $\tilde{C}_{ij} \ = C_i \wedge C_j$, the order relationship is certain. Combining the two conditions discussed above, it follows that Δ_{ij} is certain if both $\tilde{Y}_{ij}' < \tilde{C}_{2ij}$ and $\tilde{X}_{ij}' < \tilde{C}_{1ij}$ are satisfied. Hence for bivariate right censored data, the condition that the two pairs is "orderable" if $\tilde{Y}'_{ij} < \tilde{C}_{2ij}$ and $\tilde{X}'_{ij} < \tilde{C}_{1ij}$. Oakes (1986) defined $Z_{ij} = I(\tilde{X}_{ij} < \tilde{C}_{1ij}, \tilde{Y}_{ij} < \tilde{C}_{2ij})$ as the indicator of an "orderable" event. This means that Δ_{ij} can be computable if $Z_{ij} = 1$. For an AC model,

$$E[\Delta_{ij} \mid \tilde{X}_{ij}, \tilde{Y}_{ij}, Z_{ij} = 1] = \frac{\theta_\alpha(S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}{\theta_\alpha(S(\tilde{X}_{ij}, \tilde{Y}_{ij})) + 1}. \tag{3.12}$$

Under the assumption of Clayton's model,

$$E[\Delta_{ij} \mid \tilde{X}_{ij}, \tilde{Y}_{ij}, Z_{ij} = 1] = \frac{\alpha}{\alpha + 1}, \tag{3.13}$$

and the resulting estimating function, which has taken censoring into account, can be written as

$$\tilde{U}(\alpha) = \sum_{i=1}^{n} \sum_{j;j>i} \tilde{W}_\alpha(\tilde{X}_{ij}, \tilde{Y}_{ij}) \times Z_{ij} \times \left(\Delta_{ij} - \frac{\alpha}{\alpha + 1} \right), \tag{3.14}$$

where $\tilde{W}(\cdot)$ is a weight function which is chosen to improve efficiency of the resulting estimator.

For semi-competing risks data with an external censoring, X is right censored by Y or C and Y is right censored by C. The "orderable" condition becomes $\tilde{X}'_{ij} < \tilde{Y}'_{ij} < \tilde{C}_{ij}$. Fine et al. (2001) defined $D_{ij} = I(\tilde{X}_{ij} < \tilde{Y}_{ij} < \tilde{C}_{ij})$ as the indicator of orderable event. For an AC model, we have

$$E[\Delta_{ij} \mid \tilde{X}_{ij}, \tilde{Y}_{ij}, D_{ij} = 1] = \frac{\theta_\alpha(S(\tilde{X}_{ij}, \tilde{Y}_{ij}))}{\theta_\alpha(S(\tilde{X}_{ij}, \tilde{Y}_{ij})) + 1}. \tag{3.15}$$

If Clayton's model is assumed,

$$E[\Delta_{ij} \mid \tilde{X}_{ij}, \tilde{Y}_{ij}, D_{ij} = 1] = \frac{\alpha}{\alpha + 1}. \tag{3.16}$$

The resulting estimating function becomes:

$$\overline{U}(\alpha) = \sum_{i=1}^{n} \sum_{j;j>i} \overline{W}_\alpha(\tilde{X}_{ij}, \tilde{Y}_{ij}) \times D_{ij} \times \left[\Delta_{ij} - \frac{\alpha}{1 + \alpha} \right],$$

where $\overline{W}(\cdot)$ is a weight function having the effect on the efficiency as described earlier.

3.3 Estimating Functions Based on Two-by-two Tables

The paper by Day et al. (1997) and Wang (2003) show that the odds ratio of a two-by-two table contains the information of association between (X, Y) at time (x, y). In this section, we directly discuss that the developed estimating function for semi-competing risks data by Wang (2003). To simplify the discussion, we temporarily ignore the external censoring. The observed data are $\{(X_i', Y_i, \eta_i), (i = 1,.., n)\}$, where $X_i' = X_i \wedge Y_i$ and $\eta_i = I(X_i < Y_i)$. For a observed point (x, y), where $y > x$, we can construct the two-by-two table depicted in Figure 3.3.

	$X = x$	$X > x$	
$Y = y$	$N_{11}(dx, dy)$		$N_{1\bullet}(x, dy)$
$Y > y$			
	$N_{\bullet 1}(dx, y)$		$N(x,y)$

Figure 3.3 : two-by-two table at time (x, y)

The counts in the cell and the margins can be defined as

$$N_{11}(dx, dy) = \sum_{i=1}^{n} I(X_i' = x, \eta_i = 1, Y_i = y),$$

$$N_{1\bullet}(x, dy) = \sum_{i=1}^{n} I(X_i' \geq x, Y_i = y),$$

$$N_{\bullet 1}(dx, y) = \sum_{i=1}^{n} I(X_i' = x, \eta_i = 1, Y_i \geq y),$$

$$N(x, y) = \sum_{i=1}^{n} I(X_i' \geq x, Y_i \geq y).$$

Given the marginal counts, $N_{11}(dx, dy)$ follows a hyper-geometric distribution with expectation

$$E\{N_{11}(dx,dy) \mid N_{\bullet 1}(dx,y), N_{1\bullet}(x,dy), N(x,y)\}$$
$$= \frac{\theta_{\alpha}(x,y)N_{\bullet 1}(dx,y)N_{1\bullet}(x,dy)}{\theta_{\alpha}(x,y)N_{\bullet 1}(dx,y) + N(x,y) - N_{\bullet 1}(dx,y)} \tag{3.17}$$

Day et al. (1997) and Wang (2003) suggested to construct an estimating functions for α by taking the (weighted) difference between the observed count $N_{11}(dx,dy)$ and its model-based expectation $E\{N_{11}(dx,dy) \mid N_{\bullet 1}(dx,y), N_{1\bullet}(x,dy),$ $N(x,y)\}$. Under the assumption of no ties and $\theta_{\alpha}(x,y) = \alpha$, the estimating function can be expressed as

$$\breve{U}(\alpha) = \iint_{(x,y)} \breve{W}(x,y)\left[N_{11}(dx,dy) - \frac{\alpha}{\alpha + N(x,y) - 1}\right], \tag{3.18}$$

where $\breve{W}(x,y)$ is a weight function.

Chapter 4 Inference for Semi-competing Risks Data subject to Left Truncation

In this chapter, we will study a data structure which is a combination of the three data types discussed in Section 2.1. Furthermore, we will propose an inference approach to analyzing this data structure. To simplify the discussion, in Section 4.1 and 4.2 external censoring is ignored. Modification of the proposed method for censored data will be discussed in Section 4.3.

4.1 Data Description

To interpret the data structure studied in this chapter, we use the example of "diabetes diagnosis" which has been introduced in the paper by Peng et al. (2006). After the diagnosis of diabetes, a proportion of patients may suffer from some kind of morbidity, such as nephropathy or retinopathy. The relationship between morbidity and mortality is often of interest. However if researchers only include patients of diabetes who are alive at the time when the study begins, patients who die before the study time will never be included. Such a constraint of the observational scheme tends to exclude patients with shorter survival time after diagnosis. Without taking this fact into account, the subsequent analysis will be biased especially if the proportion of potential patients being excluded in the study is not low.

Using our previous notations, we observe semi-competing risks variables (X', Y, η), where Y is the time to mortality and X the time to morbidity , X' is maximum of the time to morbidity and mortality, and $\eta = I(X < Y)$. Let A be the time to the staring date of the study which is independent of (X, Y). All the three variables are measured from diagnosis of the study. Hence only those with $Y > A$ can be included in the sample. Hence the observed data are $\{(X_i', Y_i, \eta_i), (i = 1, ..., n)\}$ only if $Y_i > A_i$. We assume (X, Y) follow the Clayton model on the upper wedge,

$$\Pr(X > x, Y > y) = \left\{\Pr(X > x)^{1-\alpha} + \Pr(Y > y)^{1-\alpha} - 1\right\}^{1/1-\alpha}, \qquad (x \le y) \qquad (4.1)$$

Figure 4.1 is showed in the paper by Peng et al. (2006). Notice that only those who were alive at the beginning of the study could be included in the sample, i.e. the solid lines on Figure 4.1. Thus two (out of six) persons in the figure will be excluded to the study, i.e. the dashed lines on the figure. The study period in the diagram is long enough to observe the death events of all the subjects in the sample and hence external censoring does not exist.

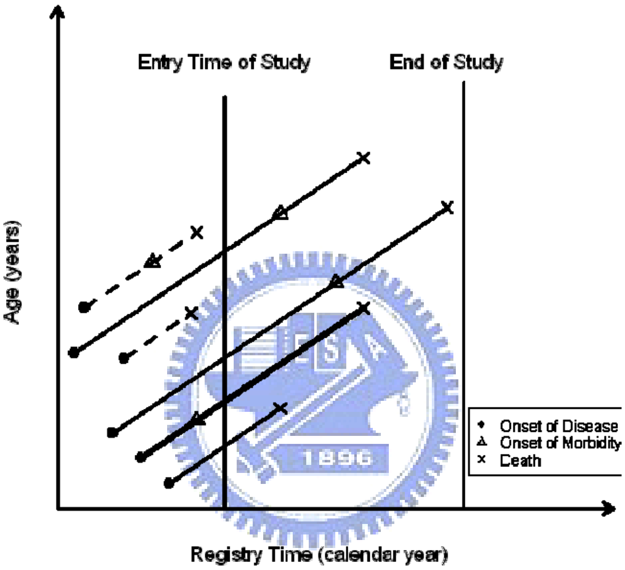


Figure 4.1: Lexis Diagram for Semi-competing Risks Data subject to Left Truncation

4.2 The Concordance Approach

As we have seen in Section 3.2, an estimating function for the association parameter can be constructed by using the information of the concordance indicator. This approach has been applied to analyze semi-competing risks data subject to left truncation.

Recall that under semi-competing risks data, we know that the information of Δ_{ij} is based on the orderable condition, $\widetilde{X}'_{ij} < \widetilde{Y}_{ij}$, which handles the censoring effect and does not involve the truncation scheme. For left truncation data, (A,Y) is observed only if $A<Y$. Consider the "comparable event" defined as $\breve{A}_{ij}<\tilde{Y}_{ij}$, where $\breve{A}_{ij}=\max(A_i,A_j)$. When this event happens, (A_i,Y_i) and (A_j,Y_j) are both located in upper wedge, $\{(a,y):0<a<y<\infty\}$. This means as long as the point $(\breve{A}_{ij},\tilde{Y}_{ij})$ is located on the upper wedge of the support of (A,Y), where $Y>A$, the (i,j) pairs are comparable. See Figure 4.2 for illustration. Combine the orderable and comparable conditions, which implies that both $\tilde{X}'_{ij}<\tilde{Y}_{ij}$ and $\breve{A}_{ij}<\tilde{Y}_{ij}$ are satisfied, define $O_{ij}=I\{(\breve{A}_{ij}\vee\tilde{X}_{ij})<\tilde{Y}_{ij}\}$ as the indicator of an "orderable" and "comparable" event.

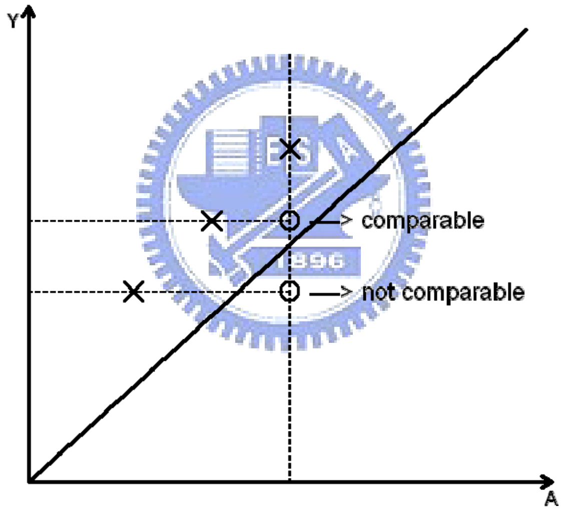


Figure 4.2: An Example of a Comparable Condition

For an AC model, it follows that

$$E[\Delta_{ij}\mid\tilde{X}_{ij},\tilde{Y}_{ij},\mathrm{O}_{ij}=1]=\frac{\theta_\alpha(S(\tilde{X}_{ij},\tilde{Y}_{ij}))}{\theta_\alpha(S(\tilde{X}_{ij},\tilde{Y}_{ij}))+1}. \tag{4.2}$$

For Clayton's model, we have

$$E[\Delta_{ij}\mid\tilde{X}_{ij},\tilde{Y}_{ij},\mathrm{O}_{ij}=1]=\frac{\alpha}{\alpha+1}, \tag{4.3}$$

Hence the resulting estimating function for the Clayton's model becomes

$$U(\alpha)=\sum_{i=1}^{n}\sum_{j;j>i}W_{\alpha}(\widetilde{X}_{ij},\widetilde{Y}_{ij})\times O_{ij}\times\left[\Delta_{ij}-\frac{\alpha}{1+\alpha}\right], \tag{4.4}$$

where $W(\cdot)$ is a weight function.

4.3 The Proposed Method Based on Two-by-Two Tables

In this section, we still use the notations of the cell and the margins of the table on Figure 3.3. Recall that in Section 2.2, Lynden-Bell's estimator uses the idea of further cutting the risk set at $Y=y$ by setting $A<y$. In Figure 4.3, the modified risk set can be written as $\{(a,y):A\le y,Y\ge y\}$.

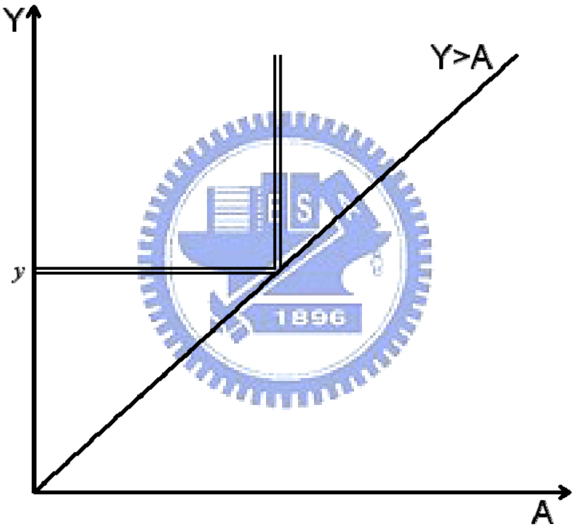


Figure 4.3 : Risk Set for Y modified for Truncation

For an observed failure point with $(X,Y)=(x,y)$, members in the original (unadjusted) risk set include those with $\{\mathrm{i}:X_i\ge x,Y_i\ge y\}$. In presence of truncation, we impose additional criteria: $\{\mathrm{i}:A_i\le y,Y_i\ge y\}$. Subjects fall in the intersection of the two sets will be included in the proposed modified risk set. The corresponding two-by-two table is given below:

	$X = x$	$X > x$	
$Y = y, A < y$	$N_{11}(dx, dy)$		$N_{1\bullet}(x, dy)$
$Y > y, A < y$	$N_{01}(dx, y)$		
	$N_{\bullet 1}(dx, y)$		$N(x,y)$

Figure 4.3 : The Proposed two-by-two Table at time (x, y)

The definitions of the cells and margins in the table are given as follows:

$$N_{11}(dx, dy) = \sum_{i=1}^{n} I(X_i' = x, \eta_i = 1, Y_i = y, A_i < y),$$

$$N_{1\bullet}(x, dy) = \sum_{i=1}^{n} I(X_i' \geq x, Y_i = y, A_i < y),$$

$$N_{\bullet 1}(dx, y) = \sum_{i=1}^{n} I(X_i' = x, \eta_i = 1, Y_i \geq y, A_i < y),$$

$$N(x, y) = \sum_{i=1}^{n} I(X_i' \geq x, Y_i' \geq y, A_i < y).$$

It follows that $N_{11}(dx, dy) \mid N_{1\bullet}(x, dy)$ follows a binomial distribution with $(N_{1\bullet}(x, dy)\ , p_1)$, where

$$p_1 = \frac{\Pr(X' = x, Y = y, A < y, \eta = 1)}{\Pr(X' \geq x, Y = y, A < y, \eta = 1)}. \tag{4.5}$$

Under the assumption that (X, Y) and A are independent, one can show that

$$\frac{\Pr(X' = x, Y = y, A < y, \eta = 1)}{\Pr(X' \geq x, Y = y, A < y, \eta = 1)} = \frac{\Pr(X = x, Y = y, A < y)}{\Pr(X \geq x, Y = y, A < y)} = \frac{\Pr(X = x, Y = y)}{\Pr(X \geq x, Y = y)}$$

Similar conclusion for the cell counts, $N_{01}(dx, y)$ follows a binomial distribution with, $(N(x, y) - N_{1\bullet}(x, dy), p_2)$ where

$$p_2 = \frac{\Pr(X = x, Y > y)}{\Pr(X \geq x, Y > y)}. \tag{4.6}$$

Hence we can find that given the margins counts, $N_{11}(dx, dy)$ still follows a hypergeometric distribution with expectation

$$E\{N_{11}(dx, dy) \mid N_{\bullet 1}(dx, y), N_{1\bullet}(x, dy), N(x, y)\}$$

$$= \frac{\theta_{\alpha}(x,y)N_{\bullet 1}(dx,y)N_{1\bullet}(x,dy)}{\theta_{\alpha}(x,y)N_{\bullet 1}(dx,y)+N(x,y)-N_{\bullet 1}(dx,y)} \tag{4.7}$$

Note that (4.4), (4.5) and (4.6) are derived in appendix. For Clayton's model with $\theta_{\alpha}(x,y)=\alpha$ and under the assumption of no tie and, the estimating function can be expressed as

$$U(\alpha)=\iint_{(x,y)} W(x,y)\left[N_{11}(dx,dy)-\frac{\alpha}{\alpha+N(x,y)-1}\right], \tag{4.8}$$

where $W(x,y)$ is a weight function.

With semi-competing risks data subject to left truncation and right censoring, the observed data are $\{(X_i',Y_i',\eta_i,\delta_i),(i=1,....,n)\}$. The definition of the cell and the margins are modified as following：

$$N_{11}(dx,dy)=\sum_{i=1}^{n} I(X_i'=x,\eta_i=1,Y_i'=y,\delta_i=1,A_i<y),$$

$$N_{1\bullet}(x,dy)=\sum_{i=1}^{n} I(X_i'\geq x,Y_i=y,\delta_i=1,A_i<y),$$

$$N_{\bullet 1}(dx,y)=\sum_{i=1}^{n} I(X_i'=x,\eta_i=1,Y_i\geq y,A_i<y),$$

$$N(x,y)=\sum_{i=1}^{n} I(X_i'\geq x,Y_i'\geq y,A_i<y).$$

The estimating function has the same form as above.

Chapter 5 Numerical Analysis

In this chapter, we evaluate the finite-sample performance of several estimators via simulation. Here we evaluate semi-competing risks data subject to left truncation. The former analysis ignores external censoring and the latter part includes censoring. As we have assumed that failure time variable (X,Y) follow the Clayton model. Here X and Y is denoted as the time to morbidity and the time to mortality, respectively. The joint distribution can be expressed as

$$\Pr(X > x, Y > y) = \left\{ [\Pr(X > x)]^{1-\alpha} + [\Pr(Y > y)]^{1-\alpha} - 1 \right\}^{1/(1-\alpha)}.$$

Let X and Y follow the exponential distribution with parameter λ_1 and λ_2, respectively. We can write

$$\begin{cases} X = -\dfrac{1}{\lambda_1}\log[1-U], \\ Y = \dfrac{1}{(\alpha-1)\cdot\lambda_2}\log[1-(1-U)^{-(\alpha-1)}+(1-U)^{-(\alpha-1)}(1-V)^{-\frac{(\alpha-1)}{\alpha}}], \end{cases} \tag{5.1}$$

where $U \sim \text{uniform}(0,1)$ and $V \sim \text{uniform}(0,1)$. In this simulation, we set $\lambda_1 = \lambda_2 = 0.5$. The association parameter α is transformed to Kendall's tau,

$$\tau = \frac{\alpha - 1}{\alpha + 1}. \tag{5.2}$$

Given a value of tau, α is produced. The truncation variable A is generated from a exponential distribution with parameter γ. The pair (X,Y) are generated subject to $Y > A$. Sample size n are chosen with 100 and 200, respectively. Each combination of α and n is simulated 1000 times.

(1) Without External Censoring

To fit the constraint of truncation in simulation, we have to decide the percentage of the original population being truncated. According to our simulation settings, we find that

$$\Pr(Y > A) = \gamma/(\lambda_2 + \gamma), \tag{5.3}$$

which means the probability of Y truncated by A is determined by λ_2 and γ. Here if we let $\Pr(Y > A) =$ 50%, we get γ =0.5. We set X' = $\min(X, Y)$ and the indicator $\eta = I(X < Y)$. The generated data are $\{(X_i', Y_i, A_i, \eta_i), (i = 1, \ldots\ldots, n)\}$ conditional on $Y_i > A_i$.

Two types of estimators are evaluated. One is based on the concordance approach:

$$U_C(\alpha) = \sum_{i=1}^{n} \sum_{j;j>i} W(\widetilde{X}_{ij}, \widetilde{Y}_{ij}) \times O_{ij} \times \left[\Delta_{ij} - \frac{\alpha}{1+\alpha} \right],$$

where $O_{ij} = I((\breve{A}_{ij} \vee \widetilde{X}_{ij}) < \widetilde{Y}_{ij})$. Here we consider two weight functions. One is $W(x, y) = 1$ and we denote the corresponding solution as $\hat{\alpha}_C$. The other weight function is

$$W(x, y) = \sum_{i=1}^{n} I(X_i' \geq x, Y_i \geq y) / n$$

and we denote the corresponding solution as $\widetilde{\alpha}_C$. The above estimating function ca be solved by the Newton-Raphson algorithm. The second method is our proposal which is based on the two-by-two table approach:

$$U_T(\alpha) = \sum_{(x,y)} W(x, y) \left[N_{11}(dx, dy) - \frac{\alpha}{\alpha + N(x, y) - 1} \right].$$

We denote the corresponding solution as $\hat{\alpha}_T$. The above estimating function is also solved by the Newton-Raphson algorithm.

The results are contained in Table 5.1 ~ Table 5.3. We see that in all the cases, the estimators of the association parameter are unbiased. Numerically the estimated variance is consistent. Especially the proposed estimator $\hat{\alpha}_T$ has smaller MSE in all the cases with different values of τ which measures the association between X and Y.

Table 5.1: Comparison of the two types of estimators for α with τ =0.75 and in absence of external censoring

	Method	n=100	n=200
Average Bias (MSE)	Concordance	0.2481 (1.8595)	0.0833 (0.7483)
	Weighted Concordance	0.2182 (1.9149)	0.0877 (0.7251)
	Two-by-Two Table	0.2248 (1.6624)	0.0879 (0.6569)

Table 5.2: Comparison of the two types of estimators for α with τ =0.5 and in absence of external censoring

	Method	n=100	n=200
Average Bias (MSE)	Concordance	0.0407 (0.2667)	0.0220 (0.1347)
	Weighted Concordance	0.0701 (0.2735)	0.0522 (0.1271)
	Two-by-Two Table	0.0564 (0.2458)	0.0400 (0.1189)

Table 5.3: Comparison of the two types of estimators for α with τ =0.25 and in absence of external censoring

	Method	n=100	n=200
Average Bias (MSE)	Concordance	0.0209 (0.0756)	0.0165 (0.0352)
	Weighted Concordance	0.0348 (0.0756)	0.0160 (0.0343)
	Two-by-Two Table	0.0296 (0.0685)	0.0343 (0.0316)

(2) With External Censoring

Let C be the censoring variable which follows a exponential distribution with the parameter μ. With censoring taking into account, the truncation criteria becomes conditional on $Y' > A$, where $Y' = \min(Y, C)$. Thus the percentage of the original population being truncated are adjusted as

$$\Pr(Y' > A) = \gamma / (\lambda_2 + \mu + \gamma), \qquad (5.4)$$

Specially the percentage of the truncated sample being censored by C can be calculated as

$$\Pr(Y > C \mid Y' > A) = \mu / (\lambda_2 + \mu). \qquad (5.5)$$

Under the above two conditions the censoring and truncated probabilities can be determined by those three parameters, $\lambda_2, \gamma,$ and $\mu.$. Here if we let $\Pr(Y' > A) =$ 50% and $\Pr(Y > C \mid Y' > A) = 80\%$, we get γ=0.625 and μ=0.125. We set $X' = \min(X, Y, C)$, the indicator $\eta = I(X < \min(Y, C))$, $Y' = \min(Y, C)$ and the indicator $\delta = I(Y < C)$. The generated data are $\{(X_i', Y_i', \eta_i, \delta_i), (i = 1, \ldots.., n)\}$ conditional on $Y_i' > A_i$.

Two types of estimators are evaluated. One is based on the concordance approach:

$$\tilde{U}_C(\alpha) = \sum_{i=1}^{n} \sum_{j;j>i} W(\tilde{X}_{ij}, \tilde{Y}_{ij}) \times Q_{ij} \times \left[\Delta_{ij} - \frac{\alpha}{1+\alpha} \right],$$

where $Q_{ij} = I((\breve{A}_{ij} \vee \tilde{X}_{ij}) < \tilde{Y}_{ij} < \tilde{C}_{ij})$. We have considered the two weight functions. One is $W(x, y) = 1$ and we denote the corresponding solution as $\hat{\alpha}_C$. The other weight function is

$$W(x, y) = \sum_{i=1}^{n} I(X_i' \geq x, Y_i' \geq y) / n$$

and we denote the corresponding solution as $\tilde{\alpha}_C$. The estimating function is solved by the Newton-Raphson algorithm. The second method is based on the two-by-two table approach:

$$\tilde{U}_T(\alpha) = \sum_{(x,y)} W(x, y) \left[N_{11}(dx, dy) - \frac{\alpha}{\alpha + N(x, y) - 1} \right].$$

We denote the corresponding solution as $\hat{\alpha}_T$. The above estimating function is also solved by the Newton-Raphson algorithm.

The results are contained in Table 5.4 ~ Table 5.6. We see that in all case, the estimators of the association parameter are still unbiased. Numerically the estimated variance is consistent. Especially the proposed estimator $\hat{\alpha}_T$ has smaller MSE in all the cases with different values of τ which measures the association between X and Y.

Table 5.4: Comparison of the two types of estimators for α with τ =0.75 and in presence of external censoring

	Method	n=100	n=200
Average Bias (MSE)	Concordance	0.1477 (2.1712)	0.1521 (1.0069)
	Weighted Concordance	0.2305 (2.3071)	0.1580 (0.9734)
	Two-by-Two Table	0.1770 (2.0012)	0.1449 (0.8757)

Table 5.5: Comparison of the two types of estimators for α with τ =0.5 and in presence of external censoring

	Method	n=100	n=200
Average Bias (MSE)	Concordance	0.0894 (0.3635)	0.0337 (0.1546)
	Weighted Concordance	0.1076 (0.3548)	0.0509 (0.1487)
	Two-by-Two Table	0.0965 (0.3250)	0.0424 (0.1361)

Table 5.6: Comparison of the two types of estimators for α with τ =0.25 and in presence of external censoring

	Method	n=100	n=200
Average Bias (MSE)	Concordance	0.0274 (0.1025)	0.0080 (0.0431)
	Weighted Concordance	0.0361 (0.0913)	0.0147 (0.0411)
	Two-by-Two Table	0.0338 (0.0865)	0.0121 (0.0379)

Chapter 6 Conclusion

In the thesis, we compare two types of inference procedures for estimating the association parameter of a copula model for semi-competing risks data subject to left truncation. If the truncation mechanism is ignored, the resulting analysis will be biased. We propose a log-rank type estimating function and find that it produces better results in simulations compared with the functions constructed based on the concordance indicators. Both methods involve deletion of some observations in the analysis to eliminate the bias due to truncation. A possible future extension is to utilize all the observations but apply a weighting approach to adjust for the sampling bias.

References

Andersen, P. K. (1988). Multistate Models in Survival Analysis: A study of Nephropathy and Mortality in Diabetes. *Statistics in Medicine*, **7**, 661-670.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application to epidemiological studies of familial tendency in chronic disease epidemiology. *Biometrika*, **65**, 141-151.

Day, R., Bryant, J. and Lefkopoulon, M. (1997). Adaptation of Biavariate Frailty Models for Prediction, with Application to Biological Markers as Prognostic Indicators. *Biometrika*, **84**, 45-56.

Fine, J. P., Jiang, H., and Chappell, R. (2001). On Semi-competing Risks Data. *Biometrika*, **88**, 907-919.

Genest, C. and Mackay, J. (1986). The Joy of Copulas: Bivariate Distributions with Uniform Marginals. *The American Statistician*, **40**, 280-283.

Jiang, H., Fine, J. P., and Chappell, R. (2005). Semiparametric Analysis of Survival Data with Left Truncation and Dependent Right Censoring. *Biometrics*, **61**, 567-575.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457-481.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis : Techniques for Censored and Truncated data*. New York: Springer-Verlag, Second Edition.

Lynden-Bell, D. (1971). A Method of Allowing for Known Observational Selection in Small Samples Applied to 3CR Quasars. *Monthly Notices of the Royal Astronomical Society*, **155**, 95-188.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London/Chapman and Hall, Second Edition.

Oakes, D. (1982) A model for Association in Bivariate Survival Data. *Journal of the Royal*

Statistical Society, Series B, **44**, 414-422.

Oakes, D. (1986). Semiparametric Inference in a Model for Association in Bivariate Survival Data. *Biometrika*, **73**, 353-361.

Oakes, D. (1989). Bivariate Survival Models induced by Frailties. *Journal of the American Statistical Association*, **84**, 487-493.

Peng, L. and Fine, J. P. (2006). Nonparametric Estimation with Left-Truncated Semi-competing Risks Data. *Biometrika*, **93**, 367-383.

Wang, W (2003). Estimating the Association Parameter for Copula Models under Dependent Censoring. *Journal of the Royal Statistical Society, Series B*, **65**, 257-273.

Appendix

To simplify the expressions, here we treat (X,Y) as discrete random variables since the probability calculations can be easily converted to the continuous case. It is obvious that

$$N_{11}(dx,dy) \mid N_{1\bullet}(x,dy) \sim BIN(N_{1\bullet}(x,dy), p_1),$$

where

$$p_1 = \frac{\Pr(X' = x, \eta = 1, Y = y, A < y)}{\Pr(X' \geq x, Y = y, A < y)},$$

and

$$N_{01}(dx,y) \mid N(x,y) - N_{1\bullet}(x,dy) \sim BIN(N(x,y) - N_{1\bullet}(x,dy), p_2),$$

where

$$p_2 = \frac{\Pr(X' = x, \eta = 1, Y \geq y, A < y)}{\Pr(X' \geq x, Y \geq y, A < y)}.$$

Now we show that p_1 and p_2 can be free of the truncation scheme. Under the assumption that (X,Y) and A are independent, one can show that

$$\begin{aligned}
p_1 &= \frac{\Pr(X' = x, \eta = 1, Y = y, A < y)}{\Pr(X' \geq x, Y = y, A < y)} = \frac{\Pr(X \wedge Y = x, \eta = 1, Y = y, A < y)}{\Pr(X \wedge Y \geq x, Y = y, A < y)} \\
&= \frac{\Pr(X = x, Y = y, A < y)}{\Pr(X \geq x, Y = y, A < y)} = \frac{\Pr(X = x, Y = y)\Pr(A < y)}{\Pr(X \geq x, Y = y)\Pr(A < y)} \\
&= \frac{\Pr(X = x, Y = y)}{\Pr(X \geq x, Y = y)},
\end{aligned}$$

and

$$\begin{aligned}
p_2 &= \frac{\Pr(X' = x, \eta = 1, Y > y, A < y)}{\Pr(X' \geq x, Y > y, A < y)} = \frac{\Pr(X \wedge Y = x, \eta = 1, Y > y, A < y)}{\Pr(X \wedge Y \geq x, Y > y, A < y)} \\
&= \frac{\Pr(X = x, Y > y, A < y)}{\Pr(X \geq x, Y > y, A < y)} = \frac{\Pr(X = x, Y > y)\Pr(A < y)}{\Pr(X \geq x, Y > y)\Pr(A < y)} \\
&= \frac{\Pr(X = x, Y > y)}{\Pr(X \geq x, Y > y)}.
\end{aligned}$$

Since given $N(dx,y) = n$ and $N_{1\bullet}(dx,y) = n_{1\bullet}$, we can know that the variable $N_{11}(dx,dy)$ is independent of $N_{01}(dx,y)$ intuitively. We have that

$$\Pr(N_{11}(dx,dy) = n_{11} \mid N_{\bullet 1}(dx,y) = n_{\bullet 1}, N(dx,y) = n, N_{1\bullet}(dx,y) = n_{1\bullet})$$

$$=\frac{\Pr\{N_{11}(dx,dy)=n_{11},N_{11}(dx,dy)+N_{01}(dx,y)=n_{\bullet 1}\}}{\Pr\{N_{11}(dx,dy)+N_{01}(dx,y)=n_{\bullet 1}\}}$$

$$=\frac{\Pr\{N_{11}(dx,dy)=n_{11},N_{01}(dx,y)=n_{\bullet 1}-n_{11}\}}{\Pr(N_{11}(dx,dy)+N_{01}(dx,y)=n_{\bullet 1}\}}$$

$$=\frac{\Pr\{N_{11}(dx,dy)=n_{11})\Pr(N_{01}(dx,y)=n_{\bullet 1}-n_{11}\}}{\Pr\{N_{11}(dx,dy)+N_{01}(dx,y)=n_{\bullet 1}\}}$$

$$=\frac{\binom{n_{1\bullet}}{n_{11}}p_1^{n_{11}}(1-p_1)^{n_{1\bullet}-n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}p_2^{n_{\bullet 1}-n_{11}}(1-p_2)^{n-n_{1\bullet}-n_{\bullet 1}+n_{11}}}{\sum\limits_{n_{11}=\max(0,N_{\bullet 1}(dx,y)-N(x,y)+N_{1\bullet}(x,dy))}^{\min(N_{\bullet 1}(dx,y),N_{1\bullet}(x,dy))}\binom{n_{1\bullet}}{n_{11}}p_1^{n_{11}}(1-p_1)^{n_{1\bullet}-n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}p_2^{n_{\bullet 1}-n_{11}}(1-p_2)^{n-n_{1\bullet}-n_{\bullet 1}+n_{11}}}$$

$$=\frac{\binom{n_{1\bullet}}{n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}\left(\frac{p_1/1-p_1}{p_2/1-p_2}\right)^{n_{11}}\left(\frac{p_2}{1-p_2}\right)^{n_{\bullet 1}}(1-p_1)^{n_{1\bullet}}(1-p_2)^{n-n_{1\bullet}}}{\sum\limits_{n_{11}=\max(0,N_{\bullet 1}(dx,y)-N(x,y)+N_{1\bullet}(x,dy))}^{\min(N_{\bullet 1}(dx,y),N_{1\bullet}(x,dy))}\binom{n_{1\bullet}}{n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}\left(\frac{p_1/1-p_1}{p_2/1-p_2}\right)^{n_{11}}\left(\frac{p_2}{1-p_2}\right)^{n_{\bullet 1}}(1-p_1)^{n_{1\bullet}}(1-p_2)^{n-n_{1\bullet}}}$$

$$=\frac{\binom{n_{1\bullet}}{n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}\left(\frac{p_1/1-p_1}{p_2/1-p_2}\right)^{n_{11}}}{\sum\limits_{n_{11}=\max(0,N_{\bullet 1}(dx,y)-N(x,y)+N_{1\bullet}(x,dy))}^{\min(N_{\bullet 1}(dx,y),N_{1\bullet}(x,dy))}\binom{n_{1\bullet}}{n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}\left(\frac{p_1/1-p_1}{p_2/1-p_2}\right)^{n_{11}}}$$

$$=\frac{\binom{n_{1\bullet}}{n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}\left(\frac{\Pr(X=x,Y=y)}{\Pr(X>x,Y=y)}\cdot\frac{\Pr(X>x,Y>y)}{\Pr(X=x,Y>y)}\right)^{n_{11}}}{\sum\limits_{n_{11}=\max(0,N_{\bullet 1}(dx,y)-N(x,y)+N_{1\bullet}(x,dy))}^{\min(N_{\bullet 1}(dx,y),N_{1\bullet}(x,dy))}\binom{n_{1\bullet}}{n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}\left(\frac{\Pr(X=x,Y=y)}{\Pr(X>x,Y=y)}\cdot\frac{\Pr(X>x,Y>y)}{\Pr(X=x,Y>y)}\right)^{n_{11}}}.$$

When (X,Y) follow the Clayton model, we can see that given the marginal counts $N_{11}(dx,dy)$ follows a hypergeometric distribution with the probability function equal to

$$\frac{\binom{n_{1\bullet}}{n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}\alpha^{n_{11}}}{\sum\limits_{n_{11}=\max(0,N_{\bullet 1}(dx,y)-N(x,y)+N_{1\bullet}(x,dy))}^{\min(N_{\bullet 1}(dx,y),N_{1\bullet}(x,dy))}\binom{n_{1\bullet}}{n_{11}}\binom{n-n_{1\bullet}}{n_{\bullet 1}-n_{11}}\alpha^{n_{11}}}.$$