國立交通大學

統計學研究所

碩士論文

家族發病時間資料之統計推論

－文獻回顧

Statistical Inference

based on Familial Age-onset Data

－A Literature Review

研 究 生： 賴信宇

指導教授： 王維菁 博士

中 華 民 國 九 十 九 年 六 月

家族發病時間資料之統計推論－文獻回顧

Statistical Inference based on Familial Age-onset Data

－A Literature Review

研究生：賴信宇	Student：Shin-Yu Lai
指導教授：王維菁 博士	Advisor：Dr. Wei-Jing Wang

國立交通大學

統計學研究所

碩士論文

A Thesis

Submitted to Institute of Statistics

College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年六月

家族發病時間資料之統計推論

－ 文獻回顧

學生： 賴信宇　　　　　　　　指導教授： 王維菁 博士

國立交通大學

統計學研究所

摘要

在本論文中，我們回顧有關分析家族發病時間的統計文獻。我們以統整的架構，說明如何探討變數間的關聯性，與建立聯合分佈的方法。討論也包含如何將解釋變數納入分析中的方法。在介紹完模式之理論意義與性質之後，我們討論幾個文獻的實例，了解統計推論所牽涉到的問題與解決的方法。

關鍵字：存活時間；發病時間；關聯模式

Statistical Inference based on Familial Age-onset Data
－ A Literature Review

Student：Shin-Yu Lai   Advisor：Dr. Weijing Wang

Institute of Statistics

National Chiao Tung University

Hsinchu, Taiwan

ABSTRACT

In the thesis, we review literature on statistical analysis of familial age-onset data. A unified framework is provided to illustrate how statistical methods are applied to investigate interesting scientific phenomenon. We discuss some important association measures and several approaches to handling correlated failure time variables. The extension to include the effect of covariates is also introduced. Finally we discuss applications of these theoretical models to biomedical data.

Keywords: Failure times; Age-at-onset; Copula model

誌　謝

2010/06/18　11:40 a.m. 口試完畢！

隨之而來的是教授們的恭喜聲。真沒想到我就這樣熬過來了...，百感交集。從開始做論文，我就在懷疑自己；投影片報完的那一刻，也覺得很不真實。總之，口試過了，開心！不過，隨著論文的付梓，代表在交大統計所兩年的求學過程即將劃上句點，這段時間以來的點點滴滴，有回憶、有不捨，回憶之事將沉澱於內心深處，使之更香醇，不捨之情將使我的人生更有勇氣與智慧。

本論文順利完成，幸蒙 王維菁教授在照顧家庭及小孩之餘，對於學生的研究方向、觀念啟迪、架構匡正、求學態度的殷切指導與逐一斧正，甚至學生的生活與家庭的關切，於此對老師獻上最深的敬意與謝意。本論文承蒙口試委員中研院黃信誠老師、清大 徐南蓉老師與交大 洪慧念老師的鼓勵與疏漏處之指正，使得本論文更臻完備，在此謹深致謝忱。

在研究所修業期間，感謝王秀瑛所長等全體老師在知識上的傳授，郭碧芬小姐在行政事務的協助。全體97級同窗兩年來的切磋討論與鼓勵，獲益匪淺。特別感謝鏡婷在許多方面的協助與意見、秋婷學姊的文獻解惑及摯友旻修的鼓勵；感謝交通大學提供優質的讀書環境及聘任助教職位，感謝這一年當中帶給我歡樂的交大102級小鬼們。還有其他對於在求學階段所有幫助我、關懷我及帶來歡笑的貴人們，由於篇幅有限而無法一一詳列，在此一併奉上最由衷之感謝。

最後，將本論文獻給我最摯愛的：祖母、父親、大姑姑、二姑姑、叔叔、大姊淑娟、二姊香青和三姊怡青等家人們，感謝大家無怨無悔的為我付出與無時無刻的關懷照顧，在精神上給予我莫大的支持，讓我能專注於課業研究中，信宇願將碩士學位之殊榮與家人共享。

以前我覺得謝天真是俗氣，但是現在我卻覺得這才是我真正的心聲。學歷的門票我拿到了，在未來人生之路，勢必有更多更艱難的挑戰；我想，我還需要很多磨練，尤其是在為中學教育付出的這條路上，我會鼓起勇氣的走下去，不讓幫助過我的人失望，期望自己也能夠當別人稱職的貴人。

謝謝，我的貴人們。

每每仰望天空，總是忍不住的向 老天爺說聲：

「謝謝！」

賴信宇 謹誌于

國立交通大學 統計學研究所

西元 2010 年孟夏風城

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Motivation and Background

Age-at-onset is an important quantitative trait for many complex diseases. It is often chosen as the variable of interest in biomedical studies for assessing the relationship between the disease and environmental or genetic factors. Familial age-onset data are useful for detecting genetic and shared environmental effects. Specifically if such effects exist, we expect that the age-onset times of family members are correlated. Understanding the pattern of association may shed light on the disease etiology and is an important step useful for further scientific investigation.

An age-onset variable measures the time to onset of the disease and hence is a lifetime variable which can be analyzed under the context of survival analysis. In the thesis, we review some literature on familial age-onset data which are applied to different biomedical areas. For example, it has been recognized that both environmental and genetic risk factors play important roles in many human diseases such as breast cancer (Hsu, 1996)[11], lung cancer (Li, 1997)[12], dementia, and heart disease. The purpose of the thesis is to provide a unified framework to illustrate how statistical methods are applied to investigate these scientific problems.

1.2 Outline of the Thesis

Let $(T_1, T_2, \cdots, T_k)$ be the age-onset variables for k members in a family. Note that T_i and T_j for $i \neq j$ are often correlated. In Chapter 2, we focus on the simplified case with k=2. Several association measures for (T_1, T_2) and different approaches to constructing the joint model are discussed. In Chapter 3, we illustrate how these ideas in Section 2.2.1 and 2.2.2 are applied to analyze familial age-onset data with mixed-effect. Copula analysis for familial age-onset data will discuss in Chapter 4. Chapter 5 contains concluding remarks.

Chapter 2 Association Measures and Models

In many biomedical applications, research interest focuses on the dependence relationship between two variables. Let T_1 and T_2 be two failure times of interest. Define the joint survival function as $S(t_1,t_2)=\Pr\left(T_1>t_1,T_2>t_2\right)$ and the corresponding density function as $\frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2}=f(t_1,t_2)$. In this chapter we review some commonly-seen association measures. Then we discuss different approaches to constructing models which describe the dependence structure.

2.1 Association Measures

2.1.1 Pearson's correlation

Pearson's correlation is perhaps the most popular measure of correlation which is defined as:

$$\rho=\frac{Cov(T_1,T_2)}{\sqrt{Var(T_1)Var(T_2)}}$$

This measure reflects the degree of linear relationship between two variables. Recall that $-1\le\rho\le1$ and its sign describes the direction of association and its value measures the degree of association. Since ρ is defined in terms of moments, it is not robust which implies that it may not be very suitable for skew distributions.

2.1.2 Kendall's tau

Let (T_{i1},T_{i2}) and (T_{j1},T_{j2}) $(i\neq j)$ be independent realizations from (T_1,T_2). The (i,j) pair is called "*concordant*" if $(T_{i1}-T_{j1})(T_{i2}-T_{j2})>0$ and "*discordant*" if $(T_{i1}-T_{j1})(T_{i2}-T_{j2})<0$. The population version of Kendall's tau is defined as the difference of concordance and discordance probabilities between pair (i,j). If T_1 and T_2 are continuous,

$$\tau = \Pr\left\{(T_{i1} - T_{j1})(T_{i2} - T_{j2}) > 0\right\} - \Pr\left\{(T_{i1} - T_{j1})(T_{i2} - T_{j2}) < 0\right\}$$

It is easy to see that $-1 \le \tau \le 1$ and, if (T_1, T_2) are independent, $\tau = 0$. Kendall's tau is known as a rank invariant measure since its value does not change as long as the marginal ranks remain the same. Estimation of τ has been discussed in Wang and Wells (2000)[18].

2.1.3 The odds ratio function

The above two measures, ρ and τ, describe global association. For failure time variables, one may be interested in more detailed information about the dependence structure. Consider the following two-by-two table:

		Member 2	
		$T_2 = t_2$	$T_2 > t_2$
Member 1	$T_1 = t_1$	a	b
	$T_1 > t_1$	c	d

Table 2.2: Two-by-two table at failure time (t_1, t_2)

If the failure times are discrete, the magnitude of association at time (t_1, t_2) can be measured by the odds ratio:

$$\frac{ad}{bc} = \frac{\Pr(T_1 > t_1, T_2 > t_2)\Pr(T_1 = t_1, T_2 = t_2)}{\Pr(T_1 = t_1, T_2 > t_2)\Pr(T_1 > t_1, T_2 = t_2)}. \tag{2.1a}$$

For continuous distributions, Oakes (1989)[15] proposed the following cross-ratio function:

$$\theta^*(t_1, t_2) = \frac{S(t_1, t_2)\left\{\partial^2 S(t_1, t_2)/\partial t_1 \partial t_2\right\}}{\left\{\partial S(t_1, t_2)/\partial t_1\right\}\left\{\partial S(t_1, t_2)/\partial t_2\right\}}. \tag{2.1b}$$

When (T_1, T_2) are independent, $\theta^*(t_1, t_2) = 1$ for all (t_1, t_2). Departure from 1 implies that association exists. This measure is useful for assessing how the level of association varies with time (t_1, t_2).

2.2 Model Construction

We discuss different approaches to constructing the joint distribution of (T_1, T_2).

2.2.1 Random effect approach

It is assumed that there exists some random effect denoted as ξ such that, given ξ, T_1 and T_2 are independent. This implies that the dependence is fully explained by ξ which is assumed to be one-dimensional to ensure identifiability. Depending on the context, ξ may represent traits which are common for the two variables.

Due to conditional independence, it follows that

$$S(t_1, t_2 \mid \xi) = S_1(t_1 \mid \xi) S_2(t_2 \mid \xi).$$

It remains to determine how ξ affects T_j. Usually a common assumption is that the random effect acts multiplicatively on the hazard, so that $\lambda_j(t) = \xi \lambda_{0j}(t)$, where $j = 1,2$ and $\lambda_{0j}(t)$ is the hazard function for the baseline group with $\xi = 1$. Accordingly $\Lambda_j(t) = \xi \Lambda_{0j}(t)$ and $S_j(t \mid \xi) = S_{0j}(t)^{\xi}$. The conditional joint survival function is then

$$S(t_1, t_2 \mid \xi) = S_{01}(t_1)^{\xi} S_{02}(t_2)^{\xi}.$$

Because ξ is not observable, to obtain the unconditional survival function, one needs to 'integrate out' ξ. Assume that ξ has the density $g(\xi)$. It follows that

$$\begin{aligned} S(t_1, t_2) &= \int_0^{\infty} S(t_1, t_2 \mid \xi) g(\xi) d\xi \\ &= \int_0^{\infty} e^{-\xi\left(\Lambda_{01}(t_1) + \Lambda_{02}(t_2)\right)} g(\xi) d\xi \\ &= E_{\xi}\left[S(t_1, t_2 \mid \xi) \right] \end{aligned} \tag{2.2}$$

which is the Laplace transform of $g(\xi)$ evaluated at $s = \Lambda_{01}(t_1) + \Lambda_{02}(t_2)$. Thus we can denote

$$S(t_1, t_2) = \pounds_g\left(\Lambda_{01}(t_1) + \Lambda_{02}(t_2)\right), \tag{2.3}$$

where $\pounds_g$ denotes the Laplace transform, $\pounds_g(s) = \int e^{-s\xi} g(\xi) d\xi$.

Example: Gamma frailty

A convenient assumption adopted in a number of literatures is that ξ has a gamma distribution. This distribution has the appropriate range $(0,\infty)$ and is mathematically tractable. The density of a gamma distribution with parameters α and λ is

$$g(\xi) = \xi^{\alpha-1} e^{-\lambda\xi} \lambda^{\alpha} \big/ \Gamma(\alpha) .$$

The mean and variance are $E(\xi) = \alpha/\lambda$ and $Var(\xi) = \alpha/\lambda^2$. It is often convenient to make the additional assumption $E(\xi) = 1$. To achieve this, we can set $\alpha = \lambda = 1/\theta$ such that $Var(\xi) = \theta$. Note that the Laplace transform of the gamma density is

$$\pounds\ (s) = \left(\frac{(1/\theta)}{(1/\theta)+s} \right)^{1/\theta} .$$

Using this result, we obtain that

$$\begin{aligned} S(t_1,t_2) &= \pounds_g\left(\Lambda_{01}(t_1) + \Lambda_{02}(t_2)\right) \\ &= \left(\frac{1/\theta}{1/\theta + \Lambda_{01}(t_1) + \Lambda_{02}(t_2)} \right)^{1/\theta} \\ &= \left(\theta\Lambda_{01}(t_1) + \theta\Lambda_{02}(t_2) + 1\right)^{-1/\theta} . \end{aligned} \qquad (2.4)$$

From the expression of the joint survival function, we can obtain

$$S_1(t_1) = S(t_1, 0) = \left(\theta\Lambda_{01}(t_1) + 1\right)^{-1/\theta} ,$$

where $\theta\Lambda_{01}(t_1) = S_1(t_1)^{-\theta} - 1$.

It follows that

$$S(t_1,t_2) = \left(S_1(t_1)^{-\theta} + S_2(t_2)^{-\theta} - 1\right)^{-1/\theta} . \qquad (2.5)$$

The above model was first introduced by Clayton (1978)[5] who derived this family based on a different approach. A useful result for the family of Gamma frailty is

$$\tau = \frac{\theta}{\theta + 2} . \qquad (2.6)$$

■

In addition to the random effect, there often exist observed covariates which also account for the existence of heterogeneity and dependence (Docrocq and Casella, 1996)[7]. Let Z_j be observed covariates for T_j. If the multiplicative effect on the hazard is still adopted, then we have

$$\lambda_j(t_j \mid Z_j, \xi) = \xi \lambda_{0j}(t_j \mid Z_j). \tag{2.7}$$

Assume that $\lambda_{0j}(t_j \mid Z_j)$ follows the Cox proportional hazards model with

$$\lambda_{0j}(t_j \mid Z_j) = \tilde{\lambda}_0(t_j) \exp\left(\beta_j^{'} Z_j\right), \tag{2.8}$$

where $\tilde{\lambda}_{0j}(t)$ is the hazard function for the baseline group with $\xi = 1$ and $Z_j = 0$. Due to conditional independence, the survival function follows that

$$S(t_1, t_2 \mid Z_1, Z_2, \xi) = S_1(t_1 \mid Z_1, \xi) S_2(t_2 \mid Z_2, \xi).$$

Then the un-conditional survival function is given by

$$S(t_1, t_2 \mid Z_1, Z_2) = \int_0^\infty S_1(t_1 \mid Z_1, \xi) S_2(t_2 \mid Z_2, \xi) g(\xi) d\xi. \tag{2.9}$$

With the gamma frailty, we have

$$S(t_1, t_2 \mid Z_1, Z_2) = \left(S_1(t_1 \mid Z_1)^{-\theta} + S_2(t_2 \mid Z_2)^{-\theta} - 1\right)^{-1/\theta}, \tag{2.10}$$

where $S_j(t_j \mid Z_j) = S_{0j}(t_j)^{\exp(\beta_j^{'} Z_j)}$ $(j = 1, 2)$.

2.2.2 Conditional approach

A conditional approach, from its literal definition, is to construct a joint model from two components: $T_1 \mid T_2$ and T_2 (or $T_2 \mid T_1$ and T_1). However such a straightforward derivation is order dependent. There exists a better way of model construction which turns out to be exchangeable in the two coordinates. Consider two conditional hazard functions:

$$\lambda_1(t_1 \mid T_2 = t_2) = \lim_{h \to 0} \frac{\Pr\left(T_1 \in [t_1, t_1 + h) \mid T_1 > t_1, T_2 = t_2\right)}{h}$$

and

$$\lambda_1(t_1 \mid T_2 > t_2) = \lim_{h\to 0} \frac{\Pr\left(T_1 \in [t_1, t_1+h) \mid T_1 > t_1, T_2 > t_2\right)}{h}.$$

Notice that

$$\theta(t_1,t_2) = \frac{\lambda_1(t_1 \mid T_2 = t_2)}{\lambda_1(t_1 \mid T_2 > t_2)} = \frac{\lambda_2(t_2 \mid T_1 = t_1)}{\lambda_2(t_2 \mid T_1 > t_1)} \qquad (2.11)$$

which does not depend on the order of conditioning. The parameter $\theta(t_1,t_2)$ measures the degree of association between T_1 and T_2 at bivariate time (t_1,t_2). Also notice that $\theta(t_1,t_2)=\theta^*(t_1,t_2)$ in (2.1).

Clayton (1978)[5] assumed that

$$\theta(t_1,t_2) = c$$

which implies that the dependence level does not varies with time. The model proposed for all t_1 and t_2

$$f(t_1,t_2)S(t_1,t_2) = c\int_{t_1}^{\infty} f(u,t_2)du \int_{t_2}^{\infty} f(t_1,v)dv.$$

From the assumption, one obtains the following second-order partial differential equation:

$$\frac{\partial^2\left(-\log S(t_1,t_2)\right)}{\partial t_1 \partial t_2} + (c-1)\frac{\partial\left(-\log S(t_1,t_2)\right)}{\partial t_1}\frac{\partial\left(-\log S(t_1,t_2)\right)}{\partial t_2} = 0.$$

Consequently, the solution of the differential equation is given by

$$S(t_1,t_2) = \left(S_1(t_1)^{-(c-1)} + S_2(t_2)^{-(c-1)} - 1\right)^{-1/(c-1)}. \qquad (2.12)$$

To include covariates, suppose that there exists covariates Z which is common to both T_1 and T_2 so that

$$\theta(t_1,t_2 \mid Z) = c(Z) = \exp(\gamma' Z), \qquad (2.13)$$

where γ is a parameter. Also there exist covariates Z_1 and Z_2 which are specific to marginal distributions of T_1 and T_2 as before. As a result, we have

$$S(t_1,t_2 \mid Z_1,Z_2) = \left(S_1(t_1 \mid Z_1)^{-(c(Z)-1)} + S_2(t_2 \mid Z_2)^{-(c(Z)-1)} - 1\right)^{-1/(c(Z)-1)}. \qquad (2.14)$$

2.2.3 Copula modeling

Copula models have been frequently adopted in the literature due to their nice properties. We first give the formal definition.

Definition of the copula function

Copula, expressed as C *is a n-dimensional function having uniform marginal distribution that satisfies the following three conditions:*

1. $C:[0,1]^n \to [0,1]$;
2. C *is a grounded and n-increasing function;*
3. C *has margins* C_i *that satisfies* $C_i(u) = C(1,\ldots,1,u,1,\ldots,1) = u$, $u \in [0,1]$. ■

Based on the above definition, it is clear that a copula function is the joint probability distribution for uniform random variables. The following theorem expands the application of copula functions.

Theorem 2.1 (Sklar's Theorem)

If $F(\cdot)$ *is an n-dimensional cdf (or survival function) with continuous margins* $F_1,\ldots,F_n$, *we can find the following unique copula representation:*

$$F(x_1,\ldots,x_n) = C\left(F_1(x_1),\ldots,F_n(x_n)\right).$$

■

That is for multivariate random variables X_j $(j=1,\ldots,n)$ with the marginal function $F_j(\cdot)$, we can construct their joint model under the copula structure.

Now we consider the special case of only two failure times. The paper by Genest and MacKay (1986)[9] derived useful properties of the bivariate family. The bivariate copula function can be written as

$$C(u_1,u_2) = \Pr\left(U_1 < u_1, U_2 < u_2\right) \quad (0 < u_j < 1) \tag{2.15}$$

where (U_1, U_2) are uniform $(0,1)$ variables marginally and $C(\cdot,\cdot):[0,1]^2 \to [0,1]$ characterizes the dependence structure. Note that if $C(u_1,u_2)=u_1u_2$ for all $0<u_j<1$ $(j=1,2)$, U_1 and U_2 are independent.

Let (T_1,T_2) be a pair of continuous failure times. In applications of lifetime data analysis, the copula structure is usually imposed on the survival function such that

$$\begin{aligned} S(t_1,t_2) &= \Pr\left(T_1>t_1, T_2>t_2\right) \\ &= \Pr\left(S_1(T_1)<S_1(t_1), S_2(T_2)<S_2(t_2)\right) \\ &= \Pr\left(U_1<S_1(t_1), U_2<S_2(t_2)\right) \\ &= C\left\{S_1(t_1), S_2(t_2)\right\}. \end{aligned}$$

where $S_j(T_j)$ $(j=1,2)$ is distributed as $Uniform(0,1)$, and

$$C(u_1,u_2)=S\left(S_1^{-1}(u_1), S_2^{-1}(u_2)\right)$$

was called as the "copula models" by Sklar (1959)[17]. Models in the copula family allow for separate investigation on the marginal behaviors and the dependence structure which is often the main interest. As mentioned earlier, when T_1 and T_2 are independent, $S(t_1,t_2)=S_1(t_1)S_2(t_2)$ and $C(u_1,u_2)=u_1u_2$. If $T_1=T_2$, $S(t_1,t_2)=S_1(t_1)\wedge S_2(t_2)$ such that $C(u_1,u_2)=u_1\wedge u_2$ which is the case of maximal dependence. The copula function $C(u_1,u_2)$ is usually parameterized as $C_\alpha(u_1,u_2)$, where the parameter α measures the association between (T_1,T_2). Note that α is related to Kendall's τ, then

$$\tau(\alpha)=4\iint C_\alpha(u_1,u_2)\frac{\partial^2 C_\alpha(u_1,u_2)}{\partial u_1 \partial u_2}du_1du_2-1. \qquad (2.16)$$

Semi-parametric inference of the copula parameter without specifying the marginal distributions has received substantial attentions in the literature.

The Archimedean copula (AC) family, which is a sub-class of the copula family, is attractive due to its nice analytical properties. For an AC model, the bivariate copula function $C_\alpha(u_1,u_2)$ can be further simplified as

$$C_{\alpha}(u_1,u_2)=\phi_{\alpha}^{-1}\{\phi_{\alpha}(u_1)+\phi_{\alpha}(u_2)\} \text{ for } u_1,u_2\in[0,1], \tag{2.17}$$

where $\phi_{\alpha}(\cdot):[0,1]\to[0,\infty]$ is a univariate function which has two continuous derivatives satisfying $\phi_{\alpha}(1)=0$, $\phi_{\alpha}{}'(t)=\dfrac{\partial\phi_{\alpha}(t)}{\partial t}<0$ and $\phi_{\alpha}{}''(t)=\dfrac{\partial^2\phi_{\alpha}(t)}{\partial t^2}>0$. AC models have the nice feature that the bivariate relationship can be summarized by the univariate function $\phi_{\alpha}(\cdot)$. Recall the cross-ratio function defined previously:

$$\theta^*(t_1,t_2)=\frac{\{\Pr(T_1>t_1,T_2>t_2)\}\{\partial^2\Pr(T_1>t_1,T_2>t_2)/\partial t_1\partial t_2\}}{\{\partial\Pr(T_1>t_1,T_2>t_2)/\partial t_1\}\{\partial\Pr(T_1>t_1,T_2>t_2)/\partial t_2\}}.$$

From Oakes (1989), an AC model has the nice property that

$$\theta^*(t_1,t_2)=\theta\{S(t_1,t_2)\},$$

$\theta(v)=-v\phi''(v)\big/\phi'(v)$ is a univariate function. A special property of the Clayton model reflects in its local odds ratio which can be expressed as $\theta^*(t_1,t_2)=\theta$ and $\theta\{S(t_1,t_2)\}$ does not depend on (t_1,t_2). In AC model, Kendall's tau has the analytic expression:

$$\tau=4\int_0^1\frac{\phi(v)}{\phi'(v)}dv+1.$$

If ϕ^{-1} is a Laplace transform of some distribution, Archimedean copula models reduce to proportional frailty models.

$$\phi_{\alpha}^{-1}(s)=L(s)=E\left[\exp(-s\xi)\right]=\int_0^{\infty}\exp(-s\xi)g(\xi)d\xi. \tag{2.18}$$

For the frailty model, the bivariate joint survival function can be expressed by integrating out the frailties with respect to the frailty density

$$\begin{aligned}S(t_1,t_2)&=\int_0^{\infty}S(t_1,t_2\mid\xi)g(\xi)d\xi\\&=E\left[\exp\left\{-\xi\left(\Lambda_{01}(t_1)+\Lambda_{02}(t_2)\right)\right\}\right],\end{aligned} \tag{2.19}$$

where Λ_{01} and Λ_{02} are the cumulative hazard functions conditional on ξ. Furthermore, the marginal survival function for each of the two imaging techniques can be obtained by having the age-onset time for the other diagnostic technique equal to zero in equation (2.19) and thus $S_j(t) = L\left\{\Lambda_{0j}(t)\right\}$. It follows that $\Lambda_{0j}(t) = L^{-1}\left(S_j(t)\right) = \phi\left(S_j(t)\right)$.

We can draw the relationship of sub-families for bivariate association model:

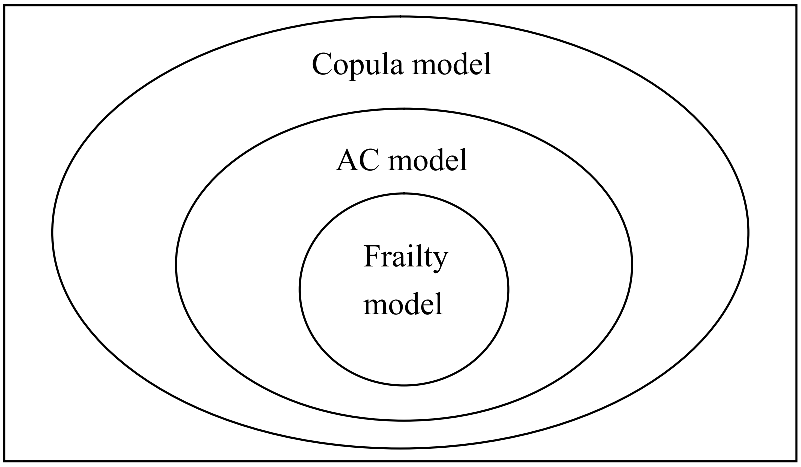

Figure 2.1: Relationship among copula models

Chapter 3 Mixed-effect Analysis for Familial Age-onset Data

In this chapter, we illustrate the application of random effect approach in analysis of familial data. Since there also exist observed covariates, the method becomes a mixed-effect approach.

3.1 Model for the Familial Study by Li and Thompson

For many common complex diseases, major susceptible genes which account for familial aggregation of the disease have been identified. For example for breast cancer, *P*53 (Malkin *et al.*, 1990)[13], *BRCA*1 (Miki *et al.*, 1994)[14] and *BRCA*2 (Wooster *et al.*, 1994)[19] are thought to cause inherited breast cancer. It has been observed that carriers with susceptible gene tend to have earlier onset ages of the disease than non-carriers.

Li and Thompson (1997)[12] modified the Cox model to analyze familial age-onset data. Let T_{ij} be the onset age for the *j*th individual in the *i*th family for $j = 1, \cdots, k_i$ and $i = 1, 2, \cdots, N$. Observed covariates for an individual include measurable genotypic information denoted as g_{ij} and other observed covariates denoted as Z_{ij}. It is important to note that $(g_{i1}, \cdots, g_{ik_i})$ are correlated which can be modeled by employing related biological knowledge. In this paper, it is assumed that a single major Mendelian diallelic locus governs disease susceptibility. Assume that there are two alleles at this locus and denote them by *A* and *a*, where *A* is the dominant disease allele with allele frequency. The genotypic covariate g_{ij} is coded as a categorical variable with three levels $(aa), (aA), (AA)$. Let the probability of $A = \Pr(A) = q$ which measures allele frequency in the population. The joint probability $\Pr(g_{i1}, \cdots, g_{ik_i})$ can be determined by Mendelian heredity law which has is the hierarchical property that the parents' information can determine the offspring's probability.

When the disease is mostly determined by the presence of the dominate gene, define

$S_{ij} = I\left[g_{ij} = Aa \text{ or } AA\right]$. The first objective is to model T_{ij} based on (S_{ij}, Z_{ij}). Li and Thompson (1997)[12] first considered the following *Cox-Gene* model:

$$\lambda_{ij}(t \mid Z_{ij}, S_{ij}) = \lambda_0(t)\exp(\beta' Z_{ij} + \mu S_{ij}) \tag{3.1a}$$

where $\lambda_0(t)$ is the hazard for the baseline group with $g_{ij} = aa$ and $Z_{ij} = 0$, β measures the effect of Z_{ij} on the hazard and μ measures the effect of having genotype "*A*" on the hazard. It is important to mention that the model in (3.1a) is a semi-parametric structure since the baseline hazard function is not specified. Compare with equation (2.8) in Section 2.2.1, (3.1a) add S_{ij} into the covariate components, where S_{ij} is the presence of the dominate gene and it is observable.

To allow for the dependence among k_i family members, the approach of random effect is also adopted. Assume that ξ_i denotes unobserved random effect within family i and it affects the hazard of T_{ij} in a multiplicative form. The following model is referred as the *Cox-Gene-Envi* model:

$$\lambda_{ij}(t \mid Z_{ij}, S_{ij}, \xi_i) = \lambda_0(t)\xi_i \exp(\beta' Z_{ij} + \mu S_{ij}). \tag{3.1b}$$

Through the influence of ξ_i on T_{ij}, $(T_{i1}, \cdots, T_{ik_i})$ become correlated and the source of association may be explained by shared environmental effects. It is further assumed that $\xi_i \overset{iid}{\sim} gamma\left(\alpha = \dfrac{1}{\theta}, \lambda = \dfrac{1}{\theta}\right)$ with the density

$$f(\xi) = \frac{\xi^{(1/\theta - 1)}\exp(-\xi/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}$$

and $E(\xi) = 1$ and $Var(\xi) = \theta$. Note that large value of θ reflects high heterogeneity in families which corresponds to larger association among family members.

The paper also discusses how to model the distribution of $\Pr(\mathbf{G}_i)$ which depends on the family structure and is calculated here under the assumptions of random mating and

Mendelian segregation which however is not our expertise. Nevertheless we summarize the basic principle. For any genotypic configuration $\mathbf{G}_i = (g_{i1},\ldots,g_{ik_i})$, one can write

$$\Pr\left(\mathbf{G}_i\right) = \prod_{\text{founder } il} \Pr\left(g_{il}\right) \prod_{\text{nonfounder } ij} \Pr\left(g_{ij} \middle| g_{m_{ij}}, g_{f_{ij}}\right) \tag{3.2}$$

where $\Pr\left(g_{il}\right)$ is the genotypic frequency and is a function of the allele frequency $\Pr(A) = q$ and $\Pr\left(g_{ij} \middle| g_{k_{ij}}, g_{f_{ij}}\right)$ is the Mendelian transmission probability given by parents. The decomposition of (3.2) depends on how members in a family are selected.

3.2 EM Algorithm for the Familial Study

Data in this example can be denoted as $(X_{ij,}\delta_{ij}, g_{ij,} Z_{ij}, \xi_i)$ for $j = 1,\ldots,k_i$ and $i = 1,\ldots,N$, where $X_{ij} = T_{ij} \wedge C_{ij} = \min(T_{ij}, C_{ij})$ and C_{ij} is the censoring time. Recall that $S_{ij} = I\left[g_{ij} = Aa \text{ or } AA\right]$ and $\Pr(\mathbf{G}_i)$ is the probability of the all members in the *i*th family with the gene frequency of A is q. Consider the full model in (3.1b). If ξ_i is observed, the contribution for *i*th family to likelihood function can be written as

$$\sum_{\mathbf{G}_i} \prod_{j=1}^{k_i} f_{Z_{ij},S_{ij}}(x_{ij} \mid \xi_i)^{\delta_{ij}} S_{Z_{ij},S_{ij}}(x_{ij} \mid \xi_i)^{1-\delta_{ij}} \Pr(\mathbf{G}_i)$$

$$= \sum_{\mathbf{G}_i} \prod_{j=1}^{k_i} \lambda_{Z_{ij},S_{ij}}(x_{ij} \mid \xi_i)^{\delta_{ij}} \exp\left\{-\Lambda_{Z_{ij},S_{ij}}(x_{ij} \mid \xi_i)\right\} \Pr(\mathbf{G}_i)$$

which is a function of $(\lambda_0(t), \beta, \mu, q)$. To include the randomness of ξ_i, the likelihood component for *i*th family becomes

$$L_i = \sum_{\mathbf{G}_i} \left[\int_{\xi_i=0}^{\infty} \prod_{j=1}^{k_i} \lambda_{S_{ij},Z_{ij}}(x_{ij} \mid \xi_i)^{\delta_{ij}} \exp\left\{-\Lambda_{S_{ij},Z_{ij}}(x_{ij} \mid \xi_i)\right\} \pi(\xi_i) d\xi_i \right] \Pr(\mathbf{G}_i)$$

which is a function of $\Theta = (\lambda_0(t), \beta, \mu, \theta, q)$. The parameter β measures the effect of Z_{ij} on the hazard; genetic effect is measured by μ; q is the frequency of genetic susceptibility, and θ measures the unobserved family-specific effect. The full likelihood function can be written as

$$\prod_{i=1}^{N}\sum_{\mathbf{G}_i}\left[\int_{\xi_i=0}^{\infty}\prod_{j=1}^{k_i}\lambda_{S_{ij},Z_{ij}}(x_{ij}\mid\xi_i)^{\delta_{ij}}\exp\left\{-\Lambda_{S_{ij},Z_{ij}}(x_{ij}\mid\xi_i)\right\}\pi(\xi_i)d\xi_i\right]\Pr(\mathbf{G}_i)$$

which however is difficult to analyze directly.

The idea of EM algorithm is applied to simplify the likelihood analysis. Temporarily assuming that $(\mathbf{G},\xi)$ are also observed. The complete data log-likelihood l_c is given by

$$\begin{aligned}
l_c &= l_c\left(\mu,\beta,\Lambda_0,\theta,q\,\middle|\,X,\delta,\mathbf{G},\xi\right)\\
&=\sum_{i=1}^{N}\sum_{j=1}^{k_i}\delta_{ij}\log\lambda_{S_{ij},Z_{ij}}(x_{ij}\mid\xi_i)-\sum_{i=1}^{N}\sum_{j=1}^{k_i}\Lambda_{g_{ij},Z_{ij}}(x_{ij}\mid\xi_i)+\sum_{i=1}^{N}\log\pi(\xi_i)+\sum_{i=1}^{N}\log\left(\sum_{G_i}\Pr(G_i)\right)\\
&=\sum_{i=1}^{N}\sum_{j=1}^{k_i}\delta_{ij}\log\left(\lambda_0(x_{ij})\xi_i\exp(\beta^{'}Z_{ij}+\mu S_{ij})\right)^{i}-\sum_{i=1}^{N}\sum_{j=1}^{k}\xi_i\Lambda_0(x_{ij})\exp(\beta^{'}Z_{ij}+\mu S_{ij})\\
&\quad+\sum_{i=1}^{N}\log\pi(\xi_i)+\sum_{i=1}^{N}\log\left(\sum_{G_i}\Pr(G_i)\right)\\
&=\sum_{i=1}^{N}\sum_{j=1}^{k_i}\delta_{ij}\log\left(\xi_i\exp(\mu S_{ij})\right)\\
&\quad-\sum_{i=1}^{N}\sum_{j=1}^{k_i}\xi_i\Lambda_0(x_{ij})\exp(\beta^{'}Z_{ij}+\mu S_{ij})+\sum_{i=1}^{N}\sum_{j=1}^{k_i}\delta_{ij}\log\left(\lambda_0(x_{ij})\exp(\beta^{'}Z_{ij})\right)\\
&\quad+\sum_{i=1}^{N}\log\pi(\xi_i)+\sum_{i=1}^{N}\log\left(\sum_{G_i}\Pr(G_i)\right)\\
&=l_1(\mu)+l_2(\mu,\beta,\Lambda_0)+l_3(\theta)+l_4(q)
\end{aligned}$$

which nicely separates the parameters.

Since $(\mathbf{G},\xi)$ are not directly observed, the E-step involves replacing $(\mathbf{G},\xi)$ by $\mathrm{E}_{\Theta}(\mathbf{G},\xi\,|\,X,\delta)$ through Monte Carlo approximation. The evaluation of expectations for the E-step requires the conditional distribution of ξ, the joint conditional distribution of $\mathbf{G}$, and the joint conditional distribution of $(\mathbf{G},\xi)$, given the observed data (X,δ). Specifically the joint conditional distribution of $(\mathbf{G},\xi)$ is given by

$$f_{\Theta}(\mathbf{G},\xi|X,\delta)=\frac{P_{\Theta}(X,\delta|\mathbf{G},\xi)P_{\Theta}(\mathbf{G})f_{\Theta}(\xi)}{\sum_{\mathbf{G}}\int_{\xi}P_{\Theta}(X,\delta|\mathbf{G},\xi)P_{\Theta}(\mathbf{G})f_{\Theta}(\xi)d\xi}. \tag{3.3}$$

It is still computationally difficult to evaluate the denominator of the above equation. The authors suggest to apply the Gibbs sampler for implementation. Let $\eta_{ij}=\xi_i\exp(\mu S_{ij})$ where $S_{ij}=\mathrm{I}(g_{ij}=Aa,AA)$ such that $\hat{\eta}_{ij}=\mathrm{E}(\eta_{ij})$. Then $\mathrm{E}(l_c)$ will be adopted as the target likelihood. To compute $\mathrm{E}(l_c)$ we need $\mathrm{E}_{\Theta}(S_{ij}|X,\delta)$, $\mathrm{E}_{\Theta}(S_{ij}\xi_i|X,\delta)$, $\mathrm{E}_{\Theta}(\xi_i)$, and $\mathrm{E}_{\Theta}(\log\xi_{\mathrm{i}})$, the expectations are computed by Monte Carlo approximation to complete the E-step of our algorithm. Finally $\mathrm{E}\left(l_4(q)\right)$ is maximized over q, $\mathrm{E}\left(l_3(\theta)\right)$ over θ, and $\mathrm{E}\left[l_1(\mu)+l_2\left(\mu,\beta,\Lambda_0(t)\right)\right]$ over $\left(\mu,\beta,\Lambda_0(t)\right)$.

3.3 Analysis for Breast Cancer Data

We list the parameter estimates about onset age of breast cancer in the paper of of Li and Thompson (1997)[12].

Parameter estimates (breast cancer)

Model	q(s.e.)	μ(s.e.)	θ(s.e.)
Cox-Gene	0.10(0.01)	3.97(0.37)	—
Cox-Gene-Envi	0.10(0.02)	2.82(0.37)	2.02(0.51)

Table 3.1: Parameter estimates based on the model Cox-Gene and Cox-Gene-Envi model

The Cox-Gene model gives a significant estimate of genetic effect. The Cox-Gene-Envi model results in significant major gene effect and also strong intra-family correlation as measured by the variance of the Gamma frailty, indicating that familial aggregation of breast cancer may be due to both gene segregation and shared familial risk.

3.4 Model for the Twin Study by Do et al.

Twin data have been used to assess genetic influences in the aetiology of complex diseases. Monozygotic (MZ) twins are genetically identical while dizygotic (DZ) twins, on the average, have half their genes in common. Therefore the association between MZ twins should be stronger than that between DZ twins. Do *et al.* (2000)[6] compared the time to menopause for two types of twins. Denote (T_{i1}, T_{i2}) as the onset ages for the twin in the ith family for $i = 1, 2, \cdots, N$. Let Z_{ij} be the observed covariates for twin j of the ith family which include the birth year and the age at menarche (both continuous), and binary variables including smoking ($0 =$ non-smokers), parity ($0 =$ fewer than 2 children) and education ($0 =$ no university education).

Modeling involves two aspects. One refers to modeling how Z_{ij} affects T_{ij} and the other is related to the dependence structure between T_{i1} and T_{i2} given Z_{i1} and Z_{i2}. In Section 2.2.1, (2.7) and (2.8) have the form that covariates Z_{ij} and the random effect component ξ_i affect T_{ij} have a proportional hazard model. And the baseline hazard rate $\lambda_0(t)$ is possibly un-specified.

For the former issue, Do *et al.* adopted a parametric approach. Specifically the marginal distribution of T_{ij} given Z_{ij} is modeled by a Weibull distribution with

$$f_{Z_{ij}}(t) = \gamma t^{\gamma-1} e^{\mu_{ij}} \exp(-e^{\mu_{ij}} t^{\gamma}), \tag{3.4}$$

where γ is the shape parameter and μ_{ij} is the scale parameter depending on Z_{ij}. An additional random-effect component is imposed on the scale parameter μ_{ij} such that

$$\log \mu_{ij} = \alpha + \beta' Z_{ij} + m_{ij}, \tag{3.5}$$

where Z_{ij} is observed covariates and m_{ij} denotes the unobserved covariate. The random effect m_{ij} affects individual heterogeneity which cannot be explained by Z_{ij} and also accounts for the dependence between T_{i1} and T_{i2}. Note that for the illustration in Section

2.2.1, the random effect component has a proportional effect on the marginal hazard functions. However in (3.4), the random effect has a proportional effect on the scale parameter μ_{ij}.

The approach of Do *et al.* (2000)[6] was motivated by the covariance component analysis proposed by Eaves *et al.* (1978)[8] who proposed to decompose the total phenotypic covariance into genetic and environmental components. Here the sources of co-variation are classified as follows:

Additive genetic factors (A) *are the effects of genes taken singly and added over multiple loci;*

Shared environmental effects (C) *are the common environmental effects.*

Different structures are imposed on m_{ij} for the two types of twins. Accordingly Do *et al.* (2000)[6] proposed the following two sub-models of m_{ij} for MZ and DZ twins. Specifically the sub-model for MZ twins is

$$\log \mu_{ij}(Z) = \alpha + \beta' Z_{ij} + m_i , \tag{3.6a}$$

where $m_{i1} = m_{i2} = m_i$ and $m_i \sim \mathrm{N}(0, \sigma_A^2 + \sigma_C^2)$. The sub-model for DZ twins is given by

$$\log \mu_{ij}(Z) = \alpha + \beta' Z_{ij} + v_i + \xi_{ij} , \tag{3.6b}$$

where $m_{ij} = v_i + \xi_{ij}$, $v_i \perp \xi_{ij}$, v_i is a random effect common for the twin pair such that

$$v_i \sim N(0, \tfrac{1}{2}\sigma_A^2 + \sigma_C^2);$$

and ξ_{ij} is a random effect for individual j such that

$$\xi_{ij} \sim N(0, \tfrac{1}{2}\sigma_A^2).$$

Notice that shared environmental effects contribute σ_C^2 to the variance of m_{ij} for both types of twins. Genetic effects contribute σ_A^2 to the variance of m_{ij} which however reveals different structures for MZ and DZ twins. Since MZ twins are genetically identical so that

σ_A^2 is common for a twin. On the other hand, DZ twins have half their genes in common while, the other half genes, are different. Thus $\frac{1}{2}\sigma_A^2$ is common for the twin pair and $\frac{1}{2}\sigma_A^2$ is left to model the variance of ξ_{ij} which measures the unobserved effect due to half different genes. In the next section, we will illustrate how the parameters are estimated which utilize Bayesian techniques.

3.5 Bayesian Inference for the Twin Study

Consider the twin data in Section 3.4. Denote onset times as $\mathbf{T}=\left(T_1,\cdots,T_K;T_{K+1},\cdots,T_N\right)$ where $T_i=(T_{i1},T_{i2})$ and the first K observations are for MZ twins and the last $N-K$ ones are for DZ twins. Model (3.4) and (3.5) gives the form of $f_{Z_{ij}}(t)$. The likelihood contribution for *i*th MZ twin is given by

$$\int_{m_i} f_{Z_{i1}}(t_{i1}\,|\,m_i)f_{Z_{i2}}(t_{i2}\,|\,m_i)\pi_1(m_i)dm_i \quad (i=1,\ldots,K), \tag{3.7a}$$

where based on (3.6a),

$$\pi_1(x)=\frac{1}{\sqrt{2\pi\Sigma_{MZ}}}\exp\left(-\frac{x^2}{2\Sigma_{MZ}}\right),$$

and $\Sigma_{MZ}=\sigma_A^2+\sigma_C^2$. Similarly the likelihood contribution for *i*th DZ twin is given by

$$\int_{\xi_{i2}}\int_{\xi_{i1}}\int_{v_i} f_{Z_{i1}}(t_{i1}\,|\,v_i,\xi_{i1})f_{Z_{i2}}(t_{i2}\,|\,v_i,\xi_{i2})\pi_2(v_i)\pi_3(\xi_{i1})\pi_3(\xi_{i2})dv_i d\xi_{i1}d\xi_{i2} \tag{3.7b}$$

where $i=K+1,\ldots,N$ and based on (3.6b),

$$\pi_2(x)=\frac{1}{\sqrt{2\pi\Sigma_{DZ}}}\exp\left(-\frac{x^2}{2\Sigma_{DZ}}\right),\quad \pi_3(x)=\frac{1}{\sqrt{2\pi\Sigma_E}}\exp\left(-\frac{x^2}{2\Sigma_E}\right),$$

and $\Sigma_{DZ}=\dfrac{1}{2}\sigma_A^2+\sigma_C^2$ and $\Sigma_E=\dfrac{1}{2}\sigma_A^2$. Knowledge of genetics is applied to construct the assumptions that $m_i\sim \mathrm{N}(0,\Sigma_{MZ})$ for MZ twins and $v_i\sim N(0,\Sigma_{DZ})$ and $\xi_{ij}\sim N(0,\Sigma_E)$ for

DZ twins. Note that $\sigma_A^2 = 2(\Sigma_{MZ} - \Sigma_{DZ})$ and $\sigma_C^2 = \Sigma_{MZ} - \sigma_A^2$. The whole likelihood function can be written as

$$f(\mathbf{T}) = \prod_{i=1}^{K} \int_{m_i} f_{Z_{i1}}(t_{i1} \mid m_i) f_{Z_{i2}}(t_{i2} \mid m_i) \pi_1(m_i) dm_i$$

$$\times \prod_{i=K+1}^{N} \int_{\xi_{i2}} \int_{\xi_{i1}} \int_{v_i} f_{Z_{i1}}(t_{i1} \mid v_i, \xi_{i1}) f_{Z_{i2}}(t_{i2} \mid v_i, \xi_{i2}) \pi_2(v_i) \pi_3(\xi_{i1}) \pi_3(\xi_{i2}) dv_i d\xi_{i1} d\xi_{i2} .$$

Direct likelihood estimation is an impossible task. Fortunately the underlying Bayesian structure is useful for implementing modern simulation techniques for parameter estimation. Denote Θ as the vector of parameters and $\mathbf{T}$ be the observed data. Denote $\pi(\Theta)$ as the prior distribution of the parameter with some random effect. The likelihood function refers to $f_{\mathbf{Z}}(\mathbf{T} \mid \Theta)$. The Posterior density is given by

$$\pi(\Theta \mid \mathbf{T}) = \frac{f_{\mathbf{Z}}(\mathbf{T} \mid \Theta)\pi(\Theta)}{f_{\mathbf{Z}}(\mathbf{T})} . \tag{3.8}$$

It is difficult to derive the distribution of $\pi(\Theta \mid \mathbf{T})$ analytically. However, the algorithm of Gibbs sampling allows one to obtain many random samples from $\pi(\Theta \mid \mathbf{T})$ without knowing its form which can be further utilized for parameter estimation. Now we apply the Gibbs sampling algorithm to the aforementioned example.

Denote $\Theta = (\gamma, \alpha, \beta, \Sigma_{MZ}, \Sigma_{DZ}, \Sigma_E)$ and $\Theta^{(0)}$ be the initial value. To obtain each component of $\Theta^{(r)}$, the estimated value in *r*th step, the following algorithm is suggested:

- draw $\gamma^{(r)}$ from $\pi(\gamma \mid \mathbf{T}, \alpha^{(r-1)}, \beta^{(r-1)}, \Sigma_{MZ}^{(r-1)}, \Sigma_{DZ}^{(r-1)}, \Sigma_E^{(r-1)})$

- draw $\alpha^{(r)}$ from $\pi(\alpha \mid \mathbf{T}, \gamma^{(r)}, \beta^{(r-1)}, \Sigma_{MZ}^{(r-1)}, \Sigma_{DZ}^{(r-1)}, \Sigma_E^{(r-1)})$

- draw $\beta^{(r)}$ from $\pi(\beta \mid \mathbf{T}, \gamma^{(r)}, \alpha^{(r)}, \Sigma_{MZ}^{(r-1)}, \Sigma_{DZ}^{(r-1)}, \Sigma_E^{(r-1)})$

- draw $\Sigma_{MZ}^{(r)}$ from $\pi(\Sigma_{MZ} \mid \mathbf{T}, \gamma^{(r)}, \alpha^{(r)}, \beta^{(r)}, \Sigma_{DZ}^{(r-1)}, \Sigma_E^{(r-1)})$

- draw $\Sigma_{DZ}^{(r)}$ from $\pi(\Sigma_{DZ} \mid \mathbf{T}, \gamma^{(r)}, \alpha^{(r)}, \beta^{(r)}, \Sigma_{MZ}^{(r)}, \Sigma_E^{(r-1)})$

- draw $\Sigma_E^{(r)}$ from $\pi(\Sigma_E \mid \mathbf{T}, \gamma^{(r)}, \alpha^{(r)}, \beta^{(r)}, \Sigma_{MZ}^{(r)}, \Sigma_{DZ}^{(r)})$.

The above successive procedure is repeated for r = 1,2,3, … until some convergence criterion is reached. Denote $\Theta^{(L)} = (\gamma^{(L)}, \alpha^{(L)}, \beta^{(L)}, \Sigma_{MZ}^{(L)}, \Sigma_{DZ}^{(L)}, \Sigma_E^{(L)})$ as the sample value obtained in the Lth step. Based on Casella and George (1992)[4], as long as L is large enough, $\Theta^{(L)}$ will be an effective random realization of Θ from $\pi(\Theta \mid \mathbf{T})$. We can generate M independent Gibbs sequences of length L, which would yield an approximate iid sample from posterior density $\pi(\Theta \mid \mathbf{T})$. Based on the sample, we can estimate Θ by the sample mean. For example to estimate β, we can take the sample average of $\beta^{(L)}$ based on M observations.

Notice that (α, β) measures the effect of observed covariates while (σ_A^2, σ_C^2) reflect genetic and environmental influences on the variation of the scale parameter. Generally speaking, larger variation corresponds to higher association. To be more specific, large variation in the population means that individuals in different families are unalike. These unobserved characteristics, on the other hand, are similar to twin members in each family which explains their association.

3.6 Data Analysis for the Twin Study

We list the partial result for time to menopause in twins for estimated regression coefficients and estimated variance components in the literature of Do *et al.* (2000)[6].

Gibbs sampling approach parameter estimates

Covariate	Coefficient β	Robust s.e.(β)	95% CI of β
(a) Mean effects			
Year of birth	− 0.029	0.0035	(-0.036,-0.023)*
Smoking	0.138	0.0788	(-0.187, 0.293)
University education	− 0.397	0.1400	(-0.676,-0.123)*
Menarche	− 0.024	0.0204	(-0.063, 0.015)
Parity	− 0.586	0.1260	(-0.830,-0.033)*

(b) Variance components			
σ_A^2	0.730	0.3290	(0.129, 1.410)*
σ_C^2	0.011	0.2400	(0.456, 0.489)

Table 3.2: Result for time to menopause in twins for estimated regression coefficients and estimated variance components

Chapter 4 Copula Analysis for Familial Age-onset Data

In the previous chapters, we have discussed how familial data are analyzed by mixed-effect models. In this chapter, we consider the application of copula models in analysis of familial age-onset data. Properties of copula models are introduced in Section 2.2.3 and here they are applied to study the mortality of twins. In genetic studies, association studies of twins provide useful information about the effect of heredity on the problem of interest. Anderson (2005)[3] analyzed the lifetime of Danish twins born in the period 1881–1930 which reveal information about how genetics affect human's lifetime. The data were divided into six groups: MZ males, DZ males, UZ (unknown zygosity) males, MZ females, DZ females and UZ females, where the twins are of the same sex and both were alive at the age of 15. Subjects were followed until 1980, and their mortality has been registered.

4.1 Copula Model for the Familial Study by Andersen

Let (T_1, T_2) be the lifetimes of twins. Let Z_j be the covariate for T_j. In the data, Z_j represents the continuous variable 'year of birth minus 1900'/100. The explanatory variable was chosen because people born later tend to live longer due to the advance of medicine and improvement of general health. The effect of Z_j on mortality is modeled through $\lambda_j(t)$ which is the marginal hazard function of T_j $(j = 1, 2)$. Specifically the Cox PH model is assumed:

$$\lambda_j(t \mid Z_j) = \lambda_0(t)\exp(\beta_j' Z_j),$$

where the baseline intensity $\lambda_0(t)$ is an unknown baseline hazard function of t and β_j is the vectors of parameters. The cumulative baseline hazard function is given by

$$\Lambda_0(t) = \int_0^t \lambda_0(s)ds.$$

In the data example, the baseline group represents the twins born in 1900.

Besides the interest in β_j which measures the marginal covariate effect on T_j, the association between the two variables is also of interest. Larger association implies that mortality is more affected by genetics. Denote $S(t_1,t_2)=\Pr(T_1>t_1,T_2>t_2)$ as the joint survival function of (T_1,T_2). Without covariates, when (T_1,T_2) come from the C_θ copula, the joint survival function can be parameterized as $S(t_1,t_2)=C_\theta\left(S_1(t_1),S_2(t_2)\right)$, where $C:[0,1]^2\to[0,1]$ characterizes the dependence structure and θ measures the degree of association. In presence of covariates, the copula model can be written as

$$S_{Z_1,Z_2}(t_1,t_2)=C_\theta\left(S_1(t_1\,|\,Z_1),S_2(t_2\,|\,Z_2)\right),$$

where under the Cox model, $S_j(t_j\,|\,Z_j)=S_{0j}(t_j)^{\exp(\beta_j' Z_j)}$ and $S_{0j}(t)=\Pr(T_j>t\,|\,Z_j)$. Note that θ does not depend on the covariates implies that the degree of association is the same for each covariate group. For an Archimedean copula model, we can write

$$S_{Z_1,Z_2}(t_1,t_2)=\phi_\theta^{-1}\left\{\phi_\theta\left(S_1(t_1\,|\,Z_1)\right)+\phi_\theta\left(S_2(t_2\,|\,Z_2)\right)\right\}.$$

Data in presence of right censoring consist of $(X_{i1},X_{i2},\delta_{i1},\delta_{i2},Z_{i1},Z_{i2})$ for $i=1,\cdots,N$, where $X_{ij}=\min(T_{ij},C_{ij})$, $\delta_{ij}=I(X_{ij}=T_{ij})$ and (C_{i1},C_{i2}) denote the censoring variables for (T_{i1},T_{i2}). Now we discuss how to estimate θ and β_j $(j=1,2)$.

4.2 Likelihood Analysis

In absence of covariates, we can write

$$\begin{aligned}\Pr(T_1>t_1,T_2>t_2)&=\Pr\left(S_1(T_1)<S_1(t_1),S_2(T_2)<S_2(t_2)\right)\\&=\Pr(U_1<u_1,U_2<u_2)\\&=C_\theta\left(u_1,u_2\right).\end{aligned}$$

In presence of covariates, let $U_{ij}=S_j(T_{ij}\,|\,Z_{ij})$, where $S_j(t\,|\,Z_j)=\Pr(T_j>t\,|\,Z_j)$. If the form of $S_j(\cdot)$ is completely specified and there is no censoring, we can construct the likelihood function of $\varphi=(\beta_1,\beta_2,\theta)$ as

$$L(\varphi)=\prod_{i=1}^{N} c_\theta(u_{i1},u_{i2}),$$

where $c_\theta(u_1,u_2)=\dfrac{\partial^2 C_\theta\left(u_1,u_2\right)}{\partial u_1 \partial u_2}$ and $u_{ij}=S_j(t_{ij}\,|\,Z_{ij})$. Now we establish the likelihood in presence of censoring. Observed data can be written as $(X_{i1},X_{i2},\delta_{i1},\delta_{i2},Z_{i1},Z_{i2})$, we can obtain: $\tilde{U}_{ij}=S_j(X_{ij}\,|\,Z_{ij})$ for $i=1,\cdots,N$ and $j=1,2$. Now we discuss when does $\tilde{U}_{ij}$ relate to U_{ij}. As $\delta_{ij}=1$, $\tilde{U}_{ij}=U_{ij}$; as $\delta_{ij}=0$, we have $X_{ij}<T_{ij}$ and $\tilde{U}_{ij}>U_{ij}$ since $S_j(\cdot)$ is a decreasing function. Accordingly, when $(\delta_{i1},\delta_{i2})=(1,1)$, $(\tilde{u}_{i1},\tilde{u}_{i2})=(u_{i1},u_{i2})$ and we assign $c_\theta\left(\tilde{u}_{i1},\tilde{u}_{i2}\right)$ to the likelihood; when $(\delta_{i1},\delta_{i2})=(1,0)$, $u_{i1}=\tilde{u}_{i1}$, $u_{i2}<\tilde{u}_{i2}$ and we assign $C_\theta^{(10)}(\tilde{u}_{i1},\tilde{u}_{i2})$ to the likelihood; when $(\delta_{i1},\delta_{i2})=(0,1)$, $u_{i1}<\tilde{u}_{i1}$, $u_{i2}=\tilde{u}_{i2}$ and we assign $C_\theta^{(01)}(\tilde{u}_{i1},\tilde{u}_{i2})$ to the likelihood; when $(\delta_{i1},\delta_{i2})=(0,0)$, $u_{i1}<\tilde{u}_{i1}$, $u_{i2}<\tilde{u}_{i2}$ and we assign $C_\theta(\tilde{u}_{i1},\tilde{u}_{i2})$ to the likelihood, where $C_\theta^{(10)}(u_1,u_2)=\dfrac{\partial}{\partial u_1}C_\theta(u_1,u_2)$ and $C_\theta^{(01)}(u_1,u_2)=\dfrac{\partial}{\partial u_2}C_\theta(u_1,u_2)$. In summary, the likelihood function can be written as

$$\prod_{i=1}^{N} c_\theta(\tilde{u}_{i1},\tilde{u}_{i2})^{\delta_{i1}\delta_{i2}} \times C_\theta^{(10)}(\tilde{u}_{i1},\tilde{u}_{i2})^{\delta_{i1}(1-\delta_{i2})} \times C_\theta^{(01)}(\tilde{u}_{i1},\tilde{u}_{i2})^{(1-\delta_{i1})\delta_{i2}} \times C_\theta(\tilde{u}_{i1},\tilde{u}_{i2})^{(1-\delta_{i1})(1-\delta_{i2})},$$

where $u_{ij}=S_{0j}(t_j)^{\exp(\beta_j' Z_j)}$ is a function of β_j. The corresponding log-likelihood function can be written as

$$\begin{aligned}\sum_{i=1}^{N} &\delta_{i1}\delta_{i2}\log c_\theta(\tilde{u}_{i1},\tilde{u}_{i2})+\delta_{i1}(1-\delta_{i2})\log C_\theta^{(10)}(\tilde{u}_{i1},\tilde{u}_{i2})\\ &+(1-\delta_{i1})\delta_{i2}\log C_\theta^{(01)}(\tilde{u}_{i1},\tilde{u}_{i2})+(1-\delta_{i1})(1-\delta_{i2})\log C_\theta(\tilde{u}_{i1},\tilde{u}_{i2}).\end{aligned}$$

Estimation of $\varphi=(\beta_1,\beta_2,\theta)$ based on the above likelihood function involves the following issues. First u_{ij} contains the nuisance function $S_{0j}(\cdot)$ and hence can be not directly observed. In additional joint estimation of $\varphi = (\beta_1, \beta_2, \theta)$ by solving the score equations simultaneous is not an easy task. Shih and Louis (1995)[16] proposed a two-stage approach. In the first stage, the marginal parameters are estimated. In the second stage, the association parameter θ is estimated after plugging in the marginal estimates obtained in the first stage. This approach was also adopted by Andersen (2005)[3] which will be discussed.

4.3 Marginal Estimation

In this section, we introduce two approaches for marginal estimation.

4.3.1 Likelihood-based approach

The first stage involves estimating pseudo-observations of (U_1, U_2) which contain the information of the nuisance functions $S_{0j}(\cdot)$ since under the Cox PH model, $U_j = S_j(T_j \mid Z_j) = S_{0j}(T_j)^{\exp(\beta_j' Z_j)}$ $(j = 1, 2)$. The marginal analysis involves estimating β_j and $S_{0j}(t_j)$. Let $t_{(1)} < t_{(2)} < \cdots < t_{(D)}$ be observed ordered event times and $Z_{(k)j}$ be the jth covariate associated with the individual whose failure time is $t_{(k)}$, $k = 1, 2, \cdots, D$. Define the risk set at time t as $\mathrm{R}(t) = \{k: \ T_{(k)} \geq t\}$ which is the set of all individuals who are still under study at a time just prior to t. The regression parameter β_j can be estimated by maximizing the following partial likelihood

$$L(\beta_j) = \prod_{k=1}^{D} \frac{\exp(\beta_j' Z_{(k)j})}{\sum_{l \in \mathrm{R}(t_k)} \exp(\beta_j' Z_{lj})} \ (j = 1, 2).$$

Let $l(\beta_j) = \ln\left[L(\beta_j) \right]$, then the maximum likelihood estimates can be obtained by maximizing $l(\beta_j)$. The score equations are solved by taking partial derivatives of $l(\beta_j)$ with respect to the β_j $(j = 1, 2)$. The estimator of the baseline function can be expressed by

Breslow's estimator:

$$\hat{S}_{0j}(t_j)=\prod_{t_k\le t}\left\{1-\frac{\delta_{kj}}{\sum\limits_{l\in \mathrm{R}(t_k)}\exp(\hat{\beta}_j^{'}Z_{lj})}\right\},\ \text{for}\ \ k=1,2,\cdots,D,\ j=1,2.$$

Then we can obtain $\hat{U}_{ij}=\hat{S}_j(T_{ij})=\hat{S}_{0j}(T_j)^{\exp(\hat{\beta}_j^{'}Z_{ij})}$ and $\breve{U}_{ij}=\hat{S}_j(X_{ij})=\hat{S}_{0j}(X_j)^{\exp(\hat{\beta}_j^{'}Z_{ij})}$ for $i=1,2,\cdots,N,\ j=1,2.$ Notice that this approach estimates β_j $(j=1,2)$ separately. Suppose the model imposes additional assumption that $\beta_1=\beta_2=\beta$, one may want to use the whole data to contain an estimator of β which is the situation considered in the next sub-section.

4.3.2 Martingale-based approach

We first introduce some basic concepts and notations. Aalen (1975)[1] analyzed survival data under the framework of counting processes and martingales. Define $\tilde{N}_{ij}(t)=I(T_{ij}\le t)$ $(t\ge 0)$ as a counting process. For right-censored data, the counting process can be modified as $N_{ij}(t)=I(X_{ij}\le t,\delta_{ij}=1)$, which is zero until the individual dies and then jumps to one. The accumulated knowledge about what has happened to all subjects in the sample up to time t is called the "filtration" of the counting process at time t which can be written as

$$\mathbf{F}_t=\sigma\left\{I(X_{ij}\le t,\delta_{ij}=1),I(X_{ij}\le t,\delta_{ij}=0)\ (i=1,\cdots,N;\ j=1,2)\right\}.$$

As time progresses, we gather more and more information so that a natural requirement is that $\mathbf{F}_s\subset\mathbf{F}_t$ for $s\le t$. We shall denote the history at an instant just prior to time t by $\mathbf{F}_{t-}$.

For a given counting process, we define $dN(t)$ as the change in the process $N(t)$ over a short time interval $[t,t+dt)$. That is $dN(t)=N\left[(t+dt)^-\right]-N(t^-)$. In the right-censored data example (assuming no ties), $dN(t)$ is one if a death occurred at t or 0, otherwise. Define $Y_{ij}(t)=I(X_{ij}\ge t)$ as the "at-risk" process for the event under censoring, then

$$E\left[dN(t)\,|\,\mathbf{F}_{t-}\right]=I(X=t,\delta=1\,|\,X\geq t)=Y(t)\lambda(t)dt\,,$$

where $\lambda(t)=\lim\limits_{\Delta\to 0}\dfrac{\Pr\left(T\in[u,u+\Delta)\middle|T\geq u\right)}{\Delta}$ is the hazard of T at time u. Accordingly one can define the martingale process $M(t)=N(t)-\Lambda(t),\ t\in[0,1]$, where $\Lambda(t)=\int_0^t Y(s)\lambda(s)ds$ is the cumulative intensity process.

Because the twins were born in the same year, they have the same covariate. It is also reasonable to assume that $\beta_1=\beta_2=\beta$ which can be estimated by solving the marginal score equation:

$$\mathrm{U}(\beta)=\sum_{i=1}^{N}\sum_{j=1}^{2}\int_0^\tau\left[Z_{ij}-\frac{s^{(1)}(\beta,u)}{s^{(0)}(\beta,u)}\right]dN_{ij}(u)\,,$$

where τ is the maximum follow-up time,

$$s^{(0)}(\beta,u)=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2}Y_{ij}(u)\exp(\beta' Z_{ij})\,,$$

$$s^{(1)}(\beta,u)=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{2}Y_{ij}(u)Z_{ij}\exp(\beta' Z_{ij})\,.$$

By solving $\mathrm{U}(\beta)=0$, we obtain $\hat{\beta}$ which can be plugged into the following equation to estimate the baseline cumulative hazard function:

$$\hat{\mathbf{\Lambda}}_0(t)=\frac{1}{N}\int_0^t\frac{dN_{..}(u)}{s^{(0)}(\hat{\beta},u)}\,,$$

where $N_{..}(t)=\sum_{i=1}^{N}\sum_{j=1}^{2}N_{ij}(t)$ and $\hat{S}_0(t)=\exp(-\hat{\mathbf{\Lambda}}_0(t))$.

4.4 Association Estimation

With the marginal estimators, we can obtain pseudo-observations $\{(\breve{U}_{i1},\breve{U}_{i2},\delta_{i1},\delta_{i2})(i=1,\ldots,N)\}$, where $\breve{U}_{ij}=\hat{S}_{0j}(X_j)^{\exp(\hat{\beta}_j' Z_{ij})}$. The likelihood of θ can be written as

$$\prod_{i=1}^{N} c_\theta(\breve{u}_{i1},\breve{u}_{i2})^{\delta_{i1}\delta_{i2}} \times C_\theta^{(10)}(\breve{u}_{i1},\breve{u}_{i2})^{\delta_{i1}(1-\delta_{i2})}$$
$$\times C_\theta^{(01)}(\breve{u}_{i1},\breve{u}_{i2})^{(1-\delta_{i1})\delta_{i2}} \times C_\theta(\breve{u}_{i1},\breve{u}_{i2})^{(1-\delta_{i1})(1-\delta_{i2})}.$$

The corresponding log-likelihood function can be written as

$$\sum_{i=1}^{N} \delta_{i1}\delta_{i2}\log c_\theta(\breve{u}_{i1},\breve{u}_{i2}) + \delta_{i1}(1-\delta_{i2})\log C_\theta^{(10)}(\breve{u}_{i1},\breve{u}_{i2})$$
$$+(1-\delta_{i1})\delta_{i2}\log C_\theta^{(01)}(\breve{u}_{i1},\breve{u}_{i2}) + (1-\delta_{i1})(1-\delta_{i2})\log C_\theta(\breve{u}_{i1},\breve{u}_{i2}).$$

This creates a pseudo score function U_θ for the parameter θ

$$\mathrm{U}_\theta(\theta) = \sum_{i=1}^{N} \frac{\partial}{\partial\theta} l\left\{\theta, \hat{\beta}_1, \hat{\beta}_2, \hat{\Lambda}_{01}, \hat{\Lambda}_{02}\right\} = 0.$$

The estimator of θ denoted as $\hat{\theta}$ can be founded by solving the above equation. For example, if the Clayton copula is assumed (i.e. $C_\theta(u_1,u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}$), the resulting log-likelihood function is given by

$$\sum_{i=1}^{N} \delta_{i1}\delta_{i2}\log\left((\theta+1)u_{i1}^{-\theta-1}u_{i2}^{-\theta-1}(u_{i1}^{-\theta}+u_{i2}^{-\theta}-1)^{(-1/\theta)-2}\right)$$
$$+\delta_{i1}(1-\delta_{i2})\log\left(u_{i1}^{-\theta-1}(u_{i1}^{-\theta}+u_{i2}^{-\theta}-1)^{(-1/\theta)-1}\right)$$
$$+(1-\delta_{i1})\delta_{i2}\log\left(u_{i2}^{-\theta-1}(u_{i1}^{-\theta}+u_{i2}^{-\theta}-1)^{(-1/\theta)-1}\right)$$
$$+(1-\delta_{i1})(1-\delta_{i2})\log\left((u_{i1}^{-\theta}+u_{i2}^{-\theta}-1)^{-1/\theta}\right).$$

4.5 Data Analysis for Mortality in Twins

The results of data analysis in Anderson's paper are summarized in Table 4.1. The combinations of sex and zygosity were analyzed separately by fitting the Clayton copula model with the covariate 'year of birth minus 1900'/100 fitted by the Cox PH model.

Sex×Zygosity	Pairs	β(s.e.)	Gamma τ
MZ males	1366	－ 1.68(0.32)	0.1957
MZ females	1450	－ 2.60(0.34)	0.1741
DZ males	2488	－ 1.62(0.22)	0.0726
DZ females	2756	－ 2.57(0.24)	0.1096
UZ males	415	0.74(0.55)	0.2369
UZ females	512	－ 0.34(0.44)	0.1731

Table 4.1: Result for mortality in twins for the gamma copula

We see that, apart from the males with unknown zygosity, there is an increased risk of death with an early year of birth. The association between MZ twins tends to be larger than that between DZ twins as one would expect.

Chapter 5 Concluding Remarks

In this thesis, we review literature on statistical inference for familial age-onset data. We first examine important association measures and different approaches to constructing correlated data. Then we discuss applications of these theoretical results to real biomedical problems.

The random effect approach provides a flexible way of modeling, which can incorporate scientific knowledge into the analysis, but the resulting inference problems are not easy to handle due to the complexity of likelihood functions. In the chosen examples, two methods are adopted, namely the EM algorithm and Gibbs sampling. The other useful approach by copula models has become increasingly popular since it allows semi-parametric inference and hence is more robust. Here the two-stage procedure is discussed. This method yields reliable estimator but may not achieve efficiency.

Reference

[1] Aalen, O. O. (1975). Statistical Inference for a Family of Counting Processes. Ph.D. Dissertation, Dept. Statistics, University of California, Berkeley.

[2] Aalen, O. O. (1994). Effects of Frailty in Survival Analysis. *Statistical Methods in Medical Research*. **3**, 227-243.

[3] Andersen, E. W. (2005). Two-Stage Estimation in Copula Models Used in Family Studies. *Lifetime Data Analysis*. **11**, 333-350.

[4] Casella, G. & George, E. I. (1992). Explaining the Gibbs Sampler**.** *The American Statistician*. **46**, 167-174.

[5] Clayton, D. G. (1978). A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biomeirika*. **65**, 141-151.

[6] Do, K-A., Broom, B. M., Kuhnert, P., Duffy, D. L., Todorov, A. A., Treloar, S. A., Martin, N. G. (2000). Genetic Analysis of the age at Menopause by Using Estimating Equations and Bayesian Random Effects Models. *Statistics in Medicine*. **19**, 1217-1235.

[7] Ducrocq, V. & Casella, G. (1996). A Bayesian Analysis of Mixed Survival Models. *Genet. Sel. Evol.* **28**, 505-529.

[8] Eaves, L. J., Last, K., Young, P. A. & Martin, N. G. (1978). Model-fitting approaches to the analysis of human behavior. *Heredity.* **41**, 249-320.

[9] Genest, C. & MacKay J. (1986). The Joy of Copulas: Bivariate Distributions with Uniform Marginals. *The American Statistician.* **40**, 280-283.

[10] Genest, C. & Rivest, L.-P. (1993). Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association*. **88**, 1034-1043.

[11] Hsu, L. & Zhao, L. P. (1996). Assessing Familial Aggregation of Age at Onset, by Using Estimating Equations, with Application to Breast Cancer. *Am. J. Hum. Genet*. **58**,

1057-1071.

[12] Li, H. & Thompson, E. (1997). Semiparametric Estimation of Major Gene and Family-Specific Random Effects for Age of Onset. *Biometrics.* **53**, 282-293.

[13] Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Jr., Nelson, C. E., Kim, D. H., Kassel, J., Gryka, M. A., Bischoff, F. Z., Tainsky, M. A. & Friend S. H. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*. **250,** 1233–1238.

[14] Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., Ding, W., Bell, R., Rosenthal, J., Hussey, C., Tran, T., McClure, M., Frye, C., Hattier, T., Phelps, R., Haugen-Strano, A., Katcher, H., Yakumo, K., Gholami, Z., Shaffer, D., Stone, S., Bayer, S., Wray, C., Bogden, R., Dayananth, P., Ward, J., Tonin, P., Narod, S., Bristow, P. K., Norris, F. H., Helvering, L., Morrison, P., Rosteck, P., Lai, M., Barrett, J. C., Lewis, C., Neuhausen, S., Cannon-Albright, L., Goldgar, D., Wiseman, R., Kamb, A. & Skolnick, M. H. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266,** 66–71.

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

[16] Shih, J. H. and Louis, T. A. (1995). Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. Biometrics, 51, 1384-1399.

[17] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*. **8**, 229-231.

[18] Wang, W. & Wells, M. T. (2000). Estimation of Kendall's tau Under Censoring. *Statistica Sinica*. **10**, 1199-1215.

[19] Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., Barfoot, R., Hamoudi, R., Patel, S., Rice, C.,

Biggs, P., Hashim, Y., Smith, A., Connor, F., Arason, A., Gudmundsson, J., Ficenec, D., Kelsell, D., Ford, D., Tonin, P., Bishop, D. T., Spurr, N. K., Ponder, B.A.J., Eeles, R., Peto, J., Devilee, P., Cornelisse, C., Lynch, H., Narod, S., Lenoir, G., Eglisson, V., Barkadottir, R. B., Easton, D. F., Bentley, D. R., Futreal, P. A., Ashworth, A., Stratton, M. R. (1995). Identification of the breast cancer susceptibility gene BRAC2. *Nature*. **378**, 789-792.