行政院國家科學委員會專題研究計畫　成果報告

針對半競爭風險資料之模式選擇方法

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-009-001-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學統計學研究所

計畫主持人：王維菁

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 11 月 6 日

國科會報告

針對半競爭風險資料之模式選擇方法

國立交通大學統計所 王維菁

中文摘要

本計劃考慮針對半競爭風險資料之雙樣本比較. 我們假設競爭風險的關聯性具有 copula model, 針對三種差異模式提出參數估計的方法. 此三種模式分別為: Cox proportional hazard model, location-shift-model, accelerated failure time model. 部分成果收錄於謝進見的碩士論文. 在論文之外, 我們考慮模式的敏感度分析與拓廣模式至 transformation model. 結果將於近期整理與投稿.

關鍵詞: 競爭風險, 雙樣本比較, 敏感度分析.

Summary

In the project, we consider two-sample comparison based on semi-competing risks data. By imposing a copula model on the dependence structure between the competing risks, we propose a unified estimating procedure to estimate the group difference parameters under three models, namely Cox PH model, the location-shift model, and the accelerated failure time model. Some results have been presented in the thesis of Hsieh (2002). Further results, including sensitivity analysis and model extension to transformation models, have also been obtained and will be submitted for publication soon.

Key words: competing risks, two-sample comparison, sensitivity analysis.

(I) Introduction, Motivation and Literature Review

Many interesting biomedical applications involve analysis of multiple endpoints. In the project, we consider a simple multistate model, also called as a “disability model” or an “illness-death” model. Here we refer *state 1* as the initial state; *state 2* as the state of disease progression (i.e. occurrence, recurrence, complications, metastases, …,etc.) and *state 3* as the absorbing state such as death. A patient may follow two different paths: initial state → progression → death or, alternatively, initial state → death. Our main objective is to compare progression of the patients in two samples.

Let X be the time to progression, Y be the time to death. Notice that X is not observable if $X > Y$ and therefore its distribution is not identifiable without further assumption. Hence statistical inference is complicated and controversial due to the problem of non-identifiability. In this project, we consider two sample comparison based on X in the identifiable region, $X \le Y$. External censoring may occur due to patient’s withdraw, loss to follow-up or the end of study. Let C be the censoring variable. In our analysis, X and Y may be correlated but it is assumed that C is independent of (X,Y) . Under right censoring, one observes the following variables:

$$\tilde{X} = X\Lambda Y\Lambda C\,,\ \ \tilde{Y} = Y\Lambda C\,,\ \ \delta^{x} = I(X \le Y\Lambda C) \ \text{ and } \ \delta^{y} = I(Y \le C)\,,$$

where $I(.)$ denotes the indicator function. The above data structure is called semi-competing risks data by Fine et al. (2001) since death is a dependent competing risk for disease progression but not vice versa.

Two-sample comparison based on X , when X is only identifiable in the half quardrant $X \le Y$, has been considered by Lin, Robins and Wei (1996) and Chang (2000). The former uses a location-shift model to describe the group effect. The latter assumes an accelerated failure time model. However both models make an implicit assumption that the dependence structure is the same for the two groups and both inference procedures produce artificial censoring. However, we feel that such an assumption sometimes may not be reasonable. In our paper we allow the dependent relationship to be different in the two groups by imposing a flexible dependence structure. The model considered here is called copula models which have attracted substantial attention in recent years due to their wide applications. One can refer to Oakes (1989) and Genest and Rivest (1993) for more thorough review of such models.

(II). Models and Proposed Methods

Because X is subject to dependent censoring, existing nonparametric inference procedures, such as the log-rank test, are not applicable. Recall that Z denotes the group indicator taking values 0 or 1. We will consider three model alternatives for measuring the group effects. The following three models will be considered:

Model 1 - the Cox model: $h_X(x|Z) = h_0(x)\exp(\theta_1 Z)$*;*

Model 2 - the location shift model: $F_X(x|Z) = F_X(x-\theta_2 Z)$*;*

Model 3 - the accelerated failure time model: $\log(X) = -\theta_3 Z + \varepsilon$,

where ε *is an error distribution.*

Under model 1, $h_X(x|Z=1) = \exp(\theta_1)h_X(x|Z=0)$ which implies that the hazard of progression in different groups are proportional. Under model 2, $P(X \geq x|Z=0) = F_X(x)$ and $P(X \geq x|Z=1) = F_X(x-\theta_2)$. Under model 3, $\log(X) = \varepsilon$ for $Z=0$ and $\log(X) = -\theta_3 + \varepsilon$ for $Z=1$. The main objective is to estimate the group difference parameters "θ_j" under the jth model assumption (j=1,2,3).

Because we allow the association between X and Y to be different for the two groups, we need to specify the underlying dependent structure. Define the upper wedge $\mathrm{P} = \{(x,y): 0 < x \leq y < \infty\}$ as the region of interest. Assume that (X,Y) follow a copula model in the region of P. Specifically their joint survival function can be expressed as

$$F(x,y) = C_\alpha\{F_X(x), F_Y(y)\} \quad (x,y) \in \mathrm{P},$$

where $C_\alpha(.,.): [0,1]^2 \to [0,1]$, $F(x,y) = \Pr(X \geq x, Y \geq y)$, $F_X(x) = \Pr(X \geq x)$ and $F_Y(y) = \Pr(Y \geq y)$. The copula class has a useful subfamily called the Archimedean copulas (AC) family. Specifically, the joint survival function of an AC model is of the form

$$F(x,y) = C_\alpha\{F_X(x), F_Y(y)\} = \Phi_\alpha^{-1}\{\Phi_\alpha[F_X(x)] + \Phi_\alpha[F_Y(y)]\} \quad (x,y) \in \mathrm{P},$$

where $\Phi_\alpha(.): [0,1] \to [0,\infty]$. The AC family contains several useful models including those proposed by Clayton, Frank, Gumbel and the log-copula model. For a copula model, the parameter α measures the level of global association and is related to Kendall's tau. Please refer to the paper by Genest and Mackay (1986) for a review.

The proposed inference procedure is summarized as follows:

a. Estimate nuisance parameters, including $G(y) = \Pr(C \geq y)$, $F(x,y)$ $(x \leq y)$, $F_Y(y)$ and α, separately. Semiparametric estimation of α has been considered by Day et al. (1997), Fine et al. (2001) and Wang (2003).

b. Estimate $F_X(x|Z=j)$ under the imposed model. Specifically straightforward calculation gives

$$F_X(x|Z=j) = \Phi_{\alpha_j}^{-1}\{\Phi_{\alpha_j}[F(x,y|Z=j)] - \Phi_{\alpha_j}[F_Y(y|Z=j)]\}.$$

The estimates obtained in step (1) can be plugged in the above equation to derive an estimator of $F_X(x)$.

Notice that the right-hand side of $F_X(x \mid Z = j)$ actually depends on y. We study two methods to remove such dependence. Method 1, called as the " diagonal approach", was proposed by Fine, Jiang and Chappell (2001), which only considers points with $x = y$. Method 2 considers taking average of $F_X(x \mid Z = j)$ at different levels of y. In simulations we found that the first method yields better results.

Now we describe the proposed method for estimating the group difference. First of all, we pool the observed times $\tilde{X}$ in the two groups and let $t_{(1)} \le t_{(2)} \le \ldots \le t_{(n)}$ be the observed ordered times in the pooled sample, where $n = n_1 + n_2$. And let $t_{(0)} = 0$. The proposed estimating equation is motivated by the following test statistic:

$$W_{Km} = \sqrt{\frac{n_1 n_2}{n}} \int_0^{t_{(n)}} W(x)[\hat{F}_X(x \mid Z = 0) - \hat{F}_X(x \mid Z = 1)]dx,$$

where $W(x) = \dfrac{n\hat{G}_1(x)\hat{G}_2(x)}{n_1\hat{G}_1(x) + n_2\hat{G}_2(x)}$ is a weight function derived in Klein and Moeschberger (p.216). The statistic W_{Km} can also be written as

$$W_{Km} = \sqrt{\frac{n_1 n_2}{n}} \int_0^{t_{(n)}} W(x)[\hat{F}_X(x \mid Z = 0) - \hat{F}_X(x \mid Z = 1)]dx,$$

where $W(x) = \dfrac{n\hat{G}_1(x)\hat{G}_2(x)}{n_1\hat{G}_1(x) + n_2\hat{G}_2(x)}$ is a weight function derived in Klein and Moeschberger (p.216). Equivalently W_{Km} can also be written as

$$W_{Km} = \sqrt{\frac{n_1 n_2}{n}} \sum_{i=1}^{n} [t_{(i)} - t_{(i-1)}] W(t_{(i)}) [\hat{F}_X(t_{(i)} \mid Z = 0) - \hat{F}_X(t_{(i)} \mid Z = 1)].$$

The parameter of group difference can be incorporated in the equation as follows.

(A). Under the Cox Model: Recall that under the Cox proportional hazard model

$$h(x \mid Z) = h_0(x)\exp(\theta_1 Z) \quad , \quad Z = 0 \text{ or } 1.$$

When θ_1 equals its true value $\tilde{\theta}_1$, it follows that

$$F_X(x \mid Z = 1) = [F_X(x \mid Z = 0)]^{\exp(\tilde{\theta}_1)}.$$

Define $g_1(x, \theta_1) = \hat{F}_X(x \mid Z = 0)^{\exp(\theta_1)} - \hat{F}_X(x \mid Z = 1)$. Hence we have

$$S_1(\theta_1)=\sqrt{\frac{n_1n_2}{n}}\sum_{i=1}^{n}[t_{(i)}-t_{(i-1)}]W(t_{(i)})g_1(t_{(i)},\theta_1)$$
$$=\sqrt{\frac{n_1n_2}{n}}\sum_{i=1}^{n}[t_{(i)}-t_{(i-1)}]W(t_{(i)})[\hat{F}_X(t_{(i)}\,|\,Z=0)^{\exp(\theta_1)}-\hat{F}_X(t_{(i)}\,|\,Z=1)].$$

The proposed estimator of θ_1 is the solution to $S_1(\theta_1)=0$, denoted as $\hat{\theta}_1$.

(B). Under the Location-Shift Model: It follows that

$$F_X(x\,|\,Z)=F_X(x-\theta_2 Z)\quad ,\quad Z=\ 0\text{ or }1.$$

Let $\tilde{\theta}_2$ be the true value of θ_2, it follows that

$$F_X(x\,|\,Z=1)=F_X(x-\tilde{\theta}_2\,|\,Z=0).$$

therefore define

$$g_2(x,\theta_2)=\hat{F}_X(x-\theta_2\,|\,Z=0)-\hat{F}_X(x\,|\,Z=1).$$

We can construct the following estimating function:

$$S_2(\theta_2)=\sqrt{\frac{n_1n_2}{n}}\sum_{i=1}^{n}[t_{(i)}-t_{(i-1)}]W(t_{(i)})g_2(t_{(i)},\theta_2)$$
$$=\sqrt{\frac{n_1n_2}{n}}\sum_{i=1}^{n}[t_{(i)}-t_{(i-1)}]W(t_{(i)})[\hat{F}_X(t_{(i)}-\theta_2\,|\,Z=0)-\hat{F}_X(t_{(i)}\,|\,Z=1)],$$

where $\hat{F}_X(x-\theta_2\,|\,Z=0)$ is the estimator $\hat{F}_X(x\,|\,Z=0)$ based on data $\{(\tilde{X}_i+\theta_2,\tilde{Y}_i+\theta_2,\delta_i^x,\delta_i^y):i=1,2,...,n_1\}$. The proposed estimator of θ_2 is the solution to $S_2(\theta_2)=0$, denoted as $\hat{\theta}_2$.

(C). Under the Accelerated failure time model: It follows that

$$\log(X)=-\theta_3 Z+\varepsilon\quad Z=\ 0\text{ or }1$$

where ε is the error distribution. Because $X=e^{-\theta_3 Z}\cdot e^{\varepsilon}$, it follows that

$$F_X(x\,|\,Z)=\Pr(X\ge x\,|\,Z)=\Pr(e^{-\theta_3 Z}\cdot e^{\varepsilon}\ge x\,|\,Z)=\Pr(e^{\varepsilon}\ge e^{\theta_3 Z}x\,|\,Z),$$

Let $\tilde{\theta}_3$ be the true value of θ_3. It is easy to see that

$$F_X(x\,|\,Z=1)=F_X(e^{\tilde{\theta}_3}x\,|\,Z=0).$$

Let $\tilde{\theta}_3$ be the true value of θ_3. It is easy to see that

$$F_X(x\,|\,Z=1)=F_X(e^{\tilde{\theta}_3}x\,|\,Z=0).$$

Define

$$g_3(x,\theta_3)=\hat{F}_X(e^{\theta_3}x\,|\,Z=0)-\hat{F}_X(x\,|\,Z=1).$$

It is natural to consider the following estimating equation:

$$\begin{aligned}S_3(\theta_3)&=\sqrt{\frac{n_1n_2}{n}}\sum_{i=1}^{n}[t_{(i)}-t_{(i-1)}]W(t_{(i)})g_3(t_{(i)},\theta_3)\\&=\sqrt{\frac{n_1n_2}{n}}\sum_{i=1}^{n}[t_{(i)}-t_{(i-1)}]W(t_{(i)})[\hat{F}_X(e^{\theta_3}t_{(i)}\,|\,Z=0)-\hat{F}_X(t_{(i)}\,|\,Z=1)],\end{aligned}$$

where $\hat{F}_X(e^{\theta_3}x\,|\,Z=0)$ is the estimator $\hat{F}_X(x\,|\,Z=0)$ based on data $\{(e^{-\theta_3}\tilde{X}_i,e^{-\theta_3}\tilde{Y}_i,\delta_i^x,\delta_i^y):i=1,2,...,n_1\}$. The proposed estimator of θ_3 is the solution to $S_3(\theta_3)=0$, denoted as $\overset{\wedge}{\theta}_3$.

(IV) Sensitivity Analysis: The objective is to assess the effect of model mis-specification for AC models. For an Archimedean copula model, we have

$$F(x,y)=\phi^{-1}\{\phi[F_X(x)]+\phi[F_Y(y)]\},$$
$$F_X(x)=\phi^{-1}\{\phi[F(x,y)]-\phi[F_Y(y)]\},$$
$$F(x,y)=\Pr(X\geq x,Y\geq y)=\xi\{F_X(x),F_Y(y)\}.$$

Define

$$F_X(x)=\phi^{-1}\{\phi[\xi(F_X(x),F_Y(y))]-\phi[F_Y(y)]\},$$
$$F_X{}^{*}(x)=\varphi^{-1}\{\varphi[\xi(F_X(x),F_Y(y))]-\varphi[F_Y(y)]\},$$

where $\phi(.)$ is the correct function and $\varphi(.)$ is the wrong function. In our paper, we compute $\max|F_X(x)-F_X^{*}(x)|$ for selected models of $\phi(.)$ and $\varphi(.)$.

(III). References:

1. Chang, S. H. (2000). A Two-Sample Comparison for Multiple Ordered Event Data. *Biometrics*, **56**, 183-189.
2. Day, R., Bryant, J. and Lefkopoulou, M. (1997). Adaptation of Bivariate Frailty Models for Prediction, with Application to Biological Markers as Prognostic Indicators. *Biometrika,* **84**, 45-56.
3. Fine, J. P., Jiang, H. and Chappell, R. (2001). On Semi-Competing Risks Data. *Biometrika,* **88, 4**, 907-919.
4. Genest, C. and Mackay, J. (1986). The Joy of Copulas: Bivariate Distributions With Uniform Marginals. *The American Statistician*, vol. **40**, No. 4.
5. Genest, C. (1987). Frank's Family of Bivariate Distributions. *Biometrika,* **74**, 3, 549-555.

6. Genest, C. and Rivest L. P. (1993). Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association*, vol. **88**, No. 423.
7. Lin, D. Y., Robins, J. M. and Wei, L. J. (1996). Comparing Two Failure Time Distribution in The Presence of Dependent Censoring. *Biometrika,* **83, 2**, 381-393.
8. Oakes, D. (1989). Bivariate Survival Models Induced by Frailties. *Journal of the American Statistical Association*. Vol. **84**, No. 406.
9. Prentice, R. L. and Cai, J. (1992). Covariance and Survivor Function Estimation Using Censored Multivariate Failure Time Data. *Biometrika*, **79**, 3, 495-512.
10. Wang, W. and Wells, M. T. (2000). Model Selection and Semiparametric Inference for Bivariate Failure-Time Data. *Journal of the American Statistical Association*. Vol. **95**, No. 449.
11. Wang (2003). Estimating the Association Parameter for Copula Models under Dependent Censoring. To appear in *JRSSB.*
12. 謝進見 (2002). “半競爭風險下雙樣本之比較” 國立交通大學碩士論文.

(IV) 自評

研究內容與計劃的原始問題一致，均是針對半競爭風險資料的分析. 然而本來打算做的模式選擇問題，因為牽涉太多不確性(需要 case-by-case 討論)，我們將方向稍作修改，討論雙樣本的比較問題. 其中對模式假設做 sensitivity analysis 的討論，亦可做為 model selection 的考慮方向之一.