東海大學統計研究所

碩士論文

Cox 比例風險模型和 Aalen 加成模型的適應性模型選擇

Adaptive model selection between
Cox proportional hazard model and Aalen additive model

研究生：蔡宗哲

指導教授：張玉媚　博士

中華民國 101 年 7 月

摘要

在醫藥的研究中，針對具共變數的右設限存活資料，Cox(1972)比例風險模型(proportional hazards model) 經常被使用來了解病人的存活情形。但實際上，具不同共變數的病人其風險不見得成比例。因此，Aalen's (1989) 模型變成為另一個被考慮的模型，在此模式，共變數對病人風險的影響是相加(additive)的效果，且允許未知的風險係數跟時間有關，如此，共變數的影響也會隨時間而改變。然而，實務上，資料背後的真實模型一般並無法得知，且 Cox 模型與 Aalen 模型在不同的情況下表現亦有所不同。如何在其中選擇一適當的模式，至今尚無相關文獻深入探討。因此，本文嘗試發展一個數據驅動(data-driven)的選模方法，即應用廣義自由度的觀念做模式的選取且透過資料擾動的技巧得到 Kullback-Leibler 損失的漸近不偏估計式，進而用相對 Kullback-Leibler 損失做為模型選擇的準則。接著透過電腦模擬的方式驗證此方法的有效性。最後再根據分析原發性膽汁性肝硬化與急性心肌梗塞的資料說明此模型選擇準則的可行性。

關鍵字：適應性模型選擇、資料擾動、Kullback-Leibler 損失、廣義自由度、Cox 比例風險模型、Aalen 加成模型

Abstract

In medical studies, Cox proportional hazards model (Cox, 1972) is the most commonly used method to analyze the survivor function of patients when the right-censored survival data are accompany with covariates which are associated with patients' physiology and conditions. However, the proportional hazards assumption is usually violated in practice. Therefore, the Aalen's additive model (Aalen, 1989) is an alternative choice under consideration. In this model, the covariates act in an additive manner on an unknown baseline hazard rate. The unknown risk coefficients in the model are allowed to be functions of time so that the effect of a covariate may vary over time. However, the two models generally perform differently under different circumstances, so that neither the Cox model nor the Aalen model is superior in all cases. How to select between them has not been explored in the literature. Therefore, we proposed a data-driven method for making a selection based on a concept of generalized degrees of freedom, resulting in an approximately unbiased estimator of the Kullback-Leibler loss via a data perturbation technique. The effectiveness of the proposed method is justified by a simulation study and also is applied to two real data sets.

Keywords:Adaptive model selection, Data perturbation, Kullback-Leibler loss, Generalized degree of freedom, Cox proportional hazards model, Aalen additive model

目錄

表目錄

圖目錄

第一章、緒論

在許多醫學及臨床的研究分析中，最重要也是最感興趣的就是比較實驗組及對照組之間存活函數的差異。但影響存活時間的通常不只是兩組投藥或治療的不同，所以當病人有重要的因素必須考慮時，我們就需要透過模型來探討共變數(covariate)和存活時間的關係。存活分析的文獻中，有許多的模型被提出，包含 Cox 比例風險模型(Cox,1972)、Aalen 加成模型(Aalen,1989)、分層 Cox 模型(Zucker,1998)、Lin and Ying's 加成模型(Lin and Ying,1994)等。每個模型的假設不盡相同，且在不同的模型下有可能導致不同的結論，所以如何在眾多模型中選取適合的模型對資料作配適就變成了一個難題。

在一般臨床的實務應用上，Cox 比例風險模型及 Aalen 加成模型最常被使用，但此兩種模型在對共變數的假設上卻有著非常大的差異。Cox 比例風險模型假設具不同共變數病人的相對風險成比例，且對風險的影響是相乘效果；Aalen 加成模型中則假設共變量的影響隨著時間而變動，且對風險的影響是相加效果。在實務上，對於資料該用哪一個模型作配適也常常引起討論，所以在這篇文章中，我們將著重在 Cox 比例風險模型和 Aalen 加成模型的選擇。

在這裡，我們考慮右設限的資料下，將使用 Kullback-Leibler 損失(KL loss)以及廣義自由度(generalized degrees of freedom, 簡稱為 GDF)作為選擇適合模型的準則。KL 損失可以被用來測量估計的精確度，所以可以用來比較不同模型下估

計的存活函數。至於廣義自由度的概念最初則是由 Ye (1998)所提出，在選擇模型上，廣義自由度的概念已經被廣泛應用到各個不同的領域中，例如：線性迴歸 Shen and Ye (2002)、廣義線性迴歸 Shen et al. (2004)、地質統計模型 Huang and Chen (2007)、曲線配適 Chen and Huang (2011)以及應用在 Cox 比例風險模型及分層 Cox 模型的選擇上(Chen and Chang,2011)。通常 KL 損失在估計上會跟模型中的未知參數有關，所以在這邊我們藉由資料擾動的技術求出 KL 損失的近似不偏估計量，此方法可以幫助我們對不同的模型做選擇。

本篇文章基本的結構如下，首先在第二章作回顧文獻，並詳細說明此篇文章中所使用的模型。第三章則會對選取模型的方法作介紹，並定義廣義自由度及估計存活函數的KL損失;且對用來估計廣義自由度的資料擾動方法也會有所說明。第四章則是透過模擬的研究驗證此方法對模型選擇的有效性。第五章是把此方法應用在實際資料分析。最後第六章是結論。

第二章、文獻探討

此章，將會介紹本篇文章所關心的 Cox 比例風險模型(Cox, 1972)及 Aalen 加成模型(Aalen, 1989)，比較兩者之間的差異性，並對兩種模型存活函數的估計方法作一些簡單說明。在右設限資料中，令T_j為第j個病人的存活時間，C_j為第j個病人的設限時間，$j = 1, \dots, n$，X_j為第j個病人的共變數。並假設給定X之下隨機變數T和C為統計獨立。我們所觀測到的右設限存活時間為$W_j = min\left(T_j, C_j\right)$，$\delta_j = I\left(T_j \le C_j\right)$為設限指標(censoring indicator)，令$t_{(1)} < t_{(2)} < \cdots < t_{(\varsigma)}$為排序的失敗時間，$R\left(t_{(j)}\right)$為時間$t_{(j)}$之下的涉險人數集合。令$d_{(j)}$為時間$t_{(j)}$之下的失敗人數，$h\left(t_j|X_j\right)$為病人$j$的風險函數，$H\left(t|X_j\right) = \sum_{t_j \le t} h\left(t_j|X_j\right)$為第$j$個病人的累積風險函數，$S\left(t|X_j\right)$為第$j$個病人的存活函數。

第一節、Cox 比例風險模型

Cox 比例風險模型為 Cox 在 1972 年所提出，為假設共變數對風險比例影響成一常數的一種模型，其模型如下：

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)exp(\beta'\boldsymbol{X})，\qquad (1)$$

其中β為未知參數向量、$\boldsymbol{X}$為共變數向量。模型中參數β的估計方法則是用最大部分概式估計(partial maximum likelihood estimator)，由 Breslow (1974)所提出的概式函數如下：

$$L(\beta) = \prod_{j=1}^{\varsigma} \frac{\exp\left[\sum_{k=1}^{p} \beta_k X_{(i)k}\right]}{\sum_{j \in R(t_j)} \exp\left[\sum_{k=1}^{p} \beta_k X_{jk}\right]}$$

對於 Cox 比例風險模型(1)的存活函數估計，則參考 Breslow 在 1975 年提出對基準累積風險函數(baseline cumulative hazard function)的估計，b為未知參數β的最大部分概似函數估計，則累積風險函數估計式如下：

$$\widehat{H}_0(t) = \sum_{t_j \le t} \frac{d_j}{\sum_{j \in R(t_j)} \exp\left(\sum_{k=1}^{p} b_k X_{jk}\right)}$$

最後可得出$\hat{S}_0(t) = \exp\left(-\widehat{H}_0(t)\right)$，則條件在共變數$\boldsymbol{X} = \boldsymbol{X_0}$之下的存活函數估計式為$\hat{S}(t|\boldsymbol{X} = \boldsymbol{X_0}) = \hat{S}_0(t)^{\exp(b'\boldsymbol{X_0})}$。

第二節、Aalen 加成模型

Aalen 加成模型為 Aalen 於 1989 年所提出，其模型如下：

$$h\left(t|\boldsymbol{Z}(t)\right) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) Z_k(t)，\qquad (2)$$

其中 $\beta_0(t)$為一任意函數，$Z_k(t)$為與時間相關之共變數，$\beta_k(t)$則為跟時間相關之參數。Aalen 加成模型(Aalen, 1989)與 Cox 比例風險模型(Cox, 1972)最大的不同點是 Aalen 加成模型假設其參數及共變數皆與時間相關。對模型中存活函數的估計可使用最小平方法，首先定義 $B_k(t) = \int_0^t \beta_k(u)du$, $k = 0,1,\cdots, p$ ，接著我們需定義一個$n \times (p+1)$維的矩陣$\boldsymbol{X}(t)$，令$X_i(t)$為$\boldsymbol{X}(t)$的第i列，如果第i個觀測對象在時間t之下仍在涉險，則$X_i(t) = Y_i(t)\left(1, Z_1(t), \ldots, Z_p(t)\right)$，否則$X_i(t)$為$(p+1)$維的零向量，其中$Y_i(t)$為第$i$個觀測對象在時間$t$之下仍涉險時等於 1，否則等於 0。

令$\boldsymbol{I}(t)$為$n \times 1$維的向量，如果第i個觀測對象在時間t死亡則$\boldsymbol{I}(t)$中第i個元素為 1，否則為 0。對$B(t) = \big(B_0(t), B_1(t), \dots, B_p(t)\big)'$的最小平方估計式如下：

$$\hat{B}(t) = \sum_{T_i \le t} [\boldsymbol{X}'(T_i)\boldsymbol{X}(T_i)]^{-1}\boldsymbol{X}'(T_i)\boldsymbol{I}(T_i)$$

估計累積風險函數則可寫成

$$\hat{H}\left[t|\boldsymbol{Z} = \big(Z_1(t), \dots, Z_p(t)\big)\right] = \hat{B}_0(t) + \sum_{k=1}^{p} \hat{B}_k(t)Z_k(t)\text{，}$$

存活函數則透過$\hat{S}(t|\boldsymbol{Z}(t)) = \exp\big(-\hat{H}(t|\boldsymbol{Z}(t))\big)$關係式求得。若$\boldsymbol{Z} = \big(Z_1, \dots, Z_p\big)'$與時間無關，則模型(2)可改寫為

$$h\big(t|\boldsymbol{Z}(t)\big) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t)Z_k\text{。} \qquad (3)$$

因此累積風險函數的估計為

$$\hat{H}\left[t|\boldsymbol{Z} = \big(Z_1(t), \dots, Z_p(t)\big)\right] = \hat{B}_0(t) + \sum_{k=1}^{p} \hat{B}_k(t)Z_k\text{，}$$

存活函數的估計則為$\hat{S}(t|\boldsymbol{Z}(t)) = \exp\big(-\hat{H}(t|\boldsymbol{Z})\big)$。

第三章、研究方法

在這一章將會介紹 Kullback-Leibler 損失和 GDF 在本篇文章的定義及計算方式，並且為了估計 GDF 而使用的資料擾動方法，在本章也會有詳細的說明。

第一節、Kullback-Leibler 損失

本節將參考將相對 KL 損失(comparative Kullback-Leibler loss)應用在右設限存活資料的文獻(Chen and Chang,2011)。令T_i為病人i的存活時間，C_i為病人i的設限時間，$\mathrm{X}_i(t)$為病人i的共變數，$i = 1, \dots, n$。我們假設在給定X(t)之下隨機變數T和C為獨立。則觀察到的右設限存活時間為$W_i = \min(T_i, C_i)$, $\delta_i = I(T_i \le C_i)$為設限指標函數(censoring indicator)，令$\mathcal{T} \equiv \{t: t_{(1)} < t_{(2)} < \cdots < t_{(\varsigma)}\}$為排序的死亡時間。令時間t時涉險的人數為$\mathrm{Y}_t = \sum_{i=1}^{n} I(W_i \ge t)$。令$\mathrm{S}(\mathrm{t}|x(t))$為共變數$x(t)$病人的存活函數且$\mathrm{t} \in \mathcal{T}$，則$\mathrm{Y}_t$服從二項分配$Bin(\mathrm{n}, \mathrm{S}(\mathrm{t}|x(t)))$。當設限發生的時候$\mathrm{S}(\mathrm{t}|x(t)) = \mathrm{S}_T(\mathrm{t}|x(t))\mathrm{S}_C(\mathrm{t}|x(t))$，其中$\mathrm{S}_T(\mathrm{t}|x(t))$和$\mathrm{S}_C(\mathrm{t}|x(t))$為具有共變數$x(t)$病人的失敗時間與設限時間的存活函數，使得$\mathrm{Y}_t$的分配不只和失敗時間有所關聯，同時也和設限時間以及共變數$x(t)$有關。根據(Chen and Chang,2011)，可將$\mathrm{t} \in \mathcal{T}$的機率密度函數表示如下：

$$p\big(y_t|\mu(t|x(t))\big) = C_{y_t}^{n}\big(S(t|x(t))\big)^{y_t}\big(1 - S(t|x(t))\big)^{n-y_t}$$

$$= exp\left\{y_t \log\left(\frac{S(t|x(t))}{1 - S(t|x(t))}\right) + n\log(1 - S(t|x(t))) + \log C_{y_t}^{n}\right\}$$

$$= exp\left\{y_t\phi\big(\mu(t|x(t))\big) + \alpha\big(\mu(t|x(t))\big) + m(y_t)\right\}$$

其中$\mu(t|x(t)) = EY_t = nS(t|x(t))$，$\phi\big(\mu(t|x(t))\big) = \log\left(\frac{\mu(t|x(t))}{n-\mu(t|x(t))}\right)$，

$\alpha\big(\mu(t|x(t))\big) = n\log(1 - \frac{\mu(t|x(t))}{n})$，$m(y_t) = \log\ C_{y_t}^n$。

設$\hat{\mu}_\gamma(t|x(t))$為在模型γ之下對$\mu(t|x(t))$的估計值，在$t \in \mathcal{T}$、$\gamma \in \Gamma \equiv \{\gamma_1, \gamma_2\} \equiv \{\text{Cox, Aalen}\}$時，對$\hat{\mu}_\gamma(t|x(t))$與$\mu(t|x(t))$之間的 KL 損失訊息定義如下：

$$\sum_{y=0}^{n} p(y|\mu(t|x(t)))\log\left\{\frac{p(y|\mu(t|x(t)))}{p(y|\hat{\mu}_\gamma(t|x(t)))}\right\}$$

$$= \left\{\phi\big(\mu(t|x(t))\big)\mu(t|x(t)) + \alpha\big(\mu(t|x(t))\big)\right\} - \left\{\phi\left(\hat{\mu}_\gamma(t|x(t))\right)\mu(t|x(t)) + \alpha\left(\hat{\mu}_\gamma(t|x(t))\right)\right\} \quad (4)$$

由於上式的左側項$\left\{\phi\big(\mu(t|x(t))\big)\mu(t|x(t)) + \alpha\big(\mu(t|x(t))\big)\right\}$與$\gamma$獨立且跟選模過程無關，所以我們捨棄左側項並定義

$$-\left\{\phi\left(\hat{\mu}_\gamma(t|x(t))\right)\mu(t|x(t)) + \alpha\left(\hat{\mu}_\gamma(t|x(t))\right) + m(y_t)\right\},$$

為$\hat{\mu}_\gamma(t|x(t))$和$\mu(t|x(t))$之間的 KL 損失訊息。接著在模型γ之下，$\hat{\mu}_\gamma(t|x(t))$與$\mu(t|x(t))$之間的相對 KL 損失訊息定義為

$$KL\big(\mu, \hat{\mu}_\gamma\big) = -\frac{1}{\varsigma}\sum_{t=t_{(1)}}^{t_{(\varsigma)}}\left\{\phi\left(\hat{\mu}_\gamma(t|x(t))\right)\mu(t|x(t)) + \alpha\left(\hat{\mu}_\gamma(t|x(t))\right) + m(y_t)\right\}$$

$$= \frac{1}{\varsigma}\left\{-\sum_{t=t_{(1)}}^{t_{(\varsigma)}}\log\ p\left(y_t|\hat{\mu}_\gamma(t|x(t))\right) + \sum_{t=t_{(1)}}^{t_{(\varsigma)}}\phi\left(\hat{\mu}_\gamma(t|x(t))\right)\left(y_t - \mu(t|x(t))\right)\right\}, \quad (5)$$

其中ς為失敗時間的個數。

第二節、廣義自由度(Generlized degree of freedom)

我們根據式(5)，選擇相對 KL 損失較小的模型，但公式(5)包含了未知函數$\mu(t|x(t))$，因此需要$KL(\mu,\hat{\mu}_\gamma)$的估計。根據 Shen et al.(2004,Section 2)，經由公式

$$E\left[KL(\mu,\hat{\mu}_\gamma)-\frac{1}{\varsigma}\left(-\sum_{t=t_{(1)}}^{t_{(\varsigma)}}\log p\left(y_t|\hat{\mu}_\gamma(t|x(t))\right)+k\right)\right]^2$$

可得到$KL(\mu,\hat{\mu}_\gamma)$的最佳估計式為

$$\frac{1}{\varsigma}\left\{-\sum_{t=t_{(1)}}^{t_{(\varsigma)}}\log p\left(y_t|\hat{\mu}_\gamma(t|x(t))\right)+D_\gamma\right\} \tag{6}$$

其中

$$D_\gamma=\sum_{t=t_{(1)}}^{t_{(\varsigma)}}E\left\{\phi\left(\hat{\mu}_\gamma(t|x(t))\right)\left(y_t-\mu(t|x(t))\right)\right\}=\sum_{t=t_{(1)}}^{t_{(\varsigma)}}cov\left(\phi\left(\hat{\mu}_\gamma(t|x(t))\right),y_t\right)$$

為在模型γ之下的廣義自由度，廣義自由度的概念最早是由 Ye (1998)所提出，當時運用在常態模型上，後來 Shen et al.(2004)進一步推廣至指數族。基本上，廣義自由度代表著存活函數估計方法的敏感度，如果存活函數的估計對於資料來說較為敏感，則廣義自由度期望會有較大的值，無論模型中參數的個數多寡。

在實務上，如果D_γ在任意模型$\gamma\in\Gamma$之下都是已知的情況，我們能夠很簡單的選擇Γ集合中使得公式(6)最小化的模型，但一般而言，D_γ會與未知參數μ相關，所以我們需要一個D_γ的不偏估計量。在下一節，我們將會介紹如何利用資料擾動的方法求得D_γ的漸近不偏估計式。

第三節、資料擾動

在此節將引用 Ye (1998)所提出的資料擾動(data perturbation,簡稱為 DP)方法來估計D_γ。首先，我們假設$\widetilde{\mathrm{Y}}_{t_{(j)}}$為在$t_{(j)} \in \mathcal{T}$服從二項分配(binomial distribution) $Bin\left(n, \tilde{S}\left(t_{(j)}\right)\right)$的隨機變數，其中$\tilde{S}\left(t_{(j)}\right) \in (0,1)$為一預先設定的值，例如可以令$\tilde{S}\left(t_{(j)}\right) = \mathrm{Y}_{t_{(j)}}/n$，$\mathrm{j} = 1, \ldots, \varsigma$，此資料擾動的想法在於透過擾動$\mathrm{Y} = \left(\mathrm{Y}_{t_{(1)}}, \ldots, \mathrm{Y}_{t_{(\varsigma)}}\right)'$得到新的$\mathrm{Y}^* = \left(Y^*_{t_{(1)}}, \ldots, Y^*_{t_{(\varsigma)}}\right)'$，進而獲得$\hat{\mu}_\gamma$對Y的敏感度，其中$Y^*_{t_{(j)}} = (1-\tau)\mathrm{Y}_{t_{(j)}} + \tau\widetilde{\mathrm{Y}}_{t_{(j)}}$，$\tau \in (0,1]$稱為資料的擾動尺度(perturbation size)。根據 Shen et al.(2004)的結果，在任意$\gamma \in \Gamma$之下利用資料擾動所得到D_γ的估計量$\widehat{D}_\gamma$為

$$\widehat{D}_\gamma = \frac{1}{\tau^2}\sum_{t=t_{(1)}}^{t_{(\varsigma)}} cov^*\left(\phi\left(\hat{\mu}^*_\gamma(t|x(t))\right), Y^*_t\right) \quad , \tag{7}$$

其中cov^*為給定Y之下的條件共變異數，$\hat{\mu}^*_\gamma(t|x(t))$則為利用擾動後資料$\mathrm{Y}^*$在模型$\gamma$之下$\mu(t|x(t))$的估計量。在 binomial 的情況，任何$\mu$在給定模型$\gamma$之下都沒有辦法得到$D_\gamma$精確的不偏估計式，但是我們可以根據估計式(7)，藉由控制樣本大小與τ得到D_γ的漸近不偏估計式(Shen et al, 2004)。從估計式(7)可以發現當τ接近 0 的時候，$\widehat{D}_\gamma$會有相當大的變異數，尤其在$\hat{\mu}_\gamma$為不連續的情況下，有可能導致模型選擇結果的不穩定。雖然在 Ye (1998)與 Huang 和 Chen (2007)的文章中有建議 DP 法對於τ的選擇並不是那麼敏感，但為求慎重，在統計模擬的過程中，還是將τ的選擇列入我們選模的因素之一。更多關於τ在適應性選模過程中的選擇在 Shen 和 Huang(2006)的文章中有更深入的探討。

為了估計$cov^*\left(\phi\left(\hat{\mu}^*_\gamma(t|x(t))\right), Y^*_t\right)$，我們需要利用蒙特卡羅法(Monte Carlo,簡

稱為 MC)，因此對應的估計式為

$$\widehat{D}_{\gamma}=\frac{1}{\tau^{2}(K-1)}\sum_{t=t_{(1)}}^{t_{(\varsigma)}}\sum_{k=1}^{K}\left(\phi\left(\hat{\mu}_{\gamma}^{*(k)}(t|x(t))\right)-\bar{\phi}\left(\hat{\mu}_{\gamma}^{*}(t|x(t))\right)\right)\left(Y_{t}^{*(k)}-\bar{Y}_{t}^{*}\right),\quad(8)$$

其中K為 DP 法的樣本數，$Y_{t}^{*(k)}=\left(Y_{t_{(1)}}^{*(k)},\dots Y_{t_{(\varsigma)}}^{*(k)}\right)'$為第$k$次擾動的樣本，

$\hat{\mu}_{\gamma}^{*(k)}(t|x(t))=n\hat{S}_{\gamma}^{*(k)}(t|x(t))$為根據$Y_{t}^{*(k)}$在模型$\gamma$之下$\mu(t|x(t))=nS(t|x(t))$的

估計量，$\bar{Y}_{t}^{*}=\frac{1}{K}\sum_{j=1}^{K}Y_{t}^{*(j)}$，$\bar{\phi}\left(\hat{\mu}_{\gamma}^{*}(t|x(t))\right)=\frac{1}{K}\sum_{j=1}^{K}\phi\left(\hat{\mu}_{\gamma}^{*(j)}(t|x(t))\right)$。

根據估計式(6)與估計式(8)，μ與$\hat{\mu}_{\gamma}$之間的相對 KL 損失的漸近不偏估計式如下：

$$\widehat{KL}\left(\mu,\hat{\mu}_{\gamma}\right)=\frac{1}{\varsigma}\left[-\sum_{t=t_{(1)}}^{t_{(\varsigma)}}\log p\left(Y_{t}\middle|\hat{\mu}_{\gamma}(t|x(t))\right)+\widehat{D}_{\gamma}\right],\quad(9)$$

我們建議的模型選擇準則為根據估計式(9)選擇模型$\hat{\gamma}\in\Gamma$滿足

$$\hat{\gamma}=\operatorname{argmin}\widehat{KL}\left(\mu,\hat{\mu}_{\gamma}\right)。\quad(10)$$

第四章、統計模擬

在這一章，我們將經由模擬探討此種選擇模型方法的有效性，為了簡單起見，考慮右設限及單一共變數的情況下，我們以 Cox 比例風險模型及 Aalen 加成模型作為模型選擇的對象，並且從不同的擾動尺度觀察對精確度(Accuracy)的影響。接著在不同的參數組合之下探討精確度的變化。

第一節、模擬步驟

在 Cox 比例風險模型的生成中，我們使用 Weibull 分配$S_0(t) = \exp(-\lambda t^{\theta})$作為基準存活函數(baseline survival function)，其中λ=0.3、θ=1.2，並且根據模型$h(t|Z) = h_0(t)\exp(\beta Z)$生成存活時間，其中$\beta$=3.5，共變數$Z$從均勻分配(0,1)中生成，設限時間則從均勻分配(0,2)中生成。

在 Aalen 加成模型的生成中，則根據模型$h(t|Z) = \beta_0(t) + \beta_1(t)Z$，其中$\beta_0(t) = \beta_0 \times t$、$\beta_1(t) = \beta_1 \times t$，$\beta_0 = 0.1$、$\beta_1 = 1.5$，共變數$Z$從均勻分配(0,1)中生成，設限時間從均勻分配(0,2)中生成。

兩個模型的資料樣本數為 100，MC 法的擾動樣本數為 100，模擬將各重覆 1000 次，擾動尺度τ則選擇(0.1, 0.3, 0.5, 0.7, 0.9)等五種尺度作探討。

第二節、模型選擇結果

藉由電腦模擬之後，我們分別將 Cox 比例風險模型與 Aalen 加成模型，在不同擾動尺度之下各模擬 1000 次的 KL 損失計算結果列於表一及表二。表內列有擾動尺度、精確度、KL 損失的平均與標準差、GDF 的平均與標準差。

從表一與表二中可看出，不管在多大的擾動尺度之下，此適應性模型選擇準則的精確度都表現得相當好。在生成模型為 Cox 比例風險模型的情況下，精確度介於 0.905 到 0.923 之間，KL 損失的平均值也明顯較 Aalen 加成模型來的小；在生成模型為 Aalen 加成模型之下，雖然 KL 損失的平均值相當接近，但精確度依然介於 0.867 到 0.881 之間。且精確度與擾動尺度之間也沒有明顯的相關性，表示擾動尺度對模型選擇結果的影響不大，這也符合 Ye (1998)與 Huang 和 Chen (2007)的文章所提出的想法。

然後我們固定擾動尺度$\tau = 0.5$，在 Cox 比例風險模型與 Aalen 加成模型不同的參數組合下進行模擬，結果分別列於表三及表四。表內列有參數組合、精確度、KL 損失的平均與標準差、GDF 的平均與標準差。其中表內的第一行為前一次模擬所使用的參數組合，將用來與其他參數組合作對照。

根據表三與表四，可以發現在不同的參數組合下的確會造成精確度的偏差，但在生成模型為 Cox 比例風險模型之下，精確度介於 0.85 至 0.922 之間；Aalen 加成模型則介於 0.74 至 0.921 之間。能夠正確選擇出生成模型的機率還是非常高的，可證明此適應性模型選擇準則的有效性。

第五章、資料分析

我們在本章將引用兩組資料做為分析的樣本依據，第一筆資料為原發性膽汁性肝硬化(primary biliary cirrhosis,簡稱為 PBC)，此筆資料首先在 Fleming 和 Harrington(1991)的文章中被描述，總共有 418 位病人在 1974 至 1984 年間進入實驗直到死亡或設限。資料中總共列舉了 17 個共變數，其中包含了年齡、性別、血清膽紅素、投藥種類、疾病的病理階段等。在此次分析中將使用其中的血清膽紅素(serumbilirubin,mg/dl)做為我們關心的共變數。血清中的膽紅素大部分來源於衰老紅細胞被破壞後產生出來的血紅蛋白衍化而成，在肝內經過葡萄糖醛酸化的叫做直接膽紅素，未在肝內經過葡萄糖醛酸化的叫做間接膽紅素，在臨床上主要用于診斷肝臟疾病，直接膽紅素升高主要常見於原發性膽汁型肝硬化、膽道瓶頸等。間接膽紅素升高則常見於溶血性疾病、新生兒黃疸或者輸血錯誤等。

第二筆資料則是被稱做 TRACE 研究群的數據集，此筆資料在 Jensen et al.(1997)被提出探討關於急性心肌梗塞(acute myocardial infarction)的風險因素，原始資料約有 6600 個病人，我們使用的數據集為原始資料隨機抽取的子集合，總共有 1000 位病人，資料內包含了性別、年齡、有無糖尿病、心室顫動及心功能衰竭等 6 個共變數。此次將選擇有無糖尿病(diabetes,1:present, 0:absent)做為共變數來進行分析。兩筆資料的分析結果將列於表五及表六。

依照表五所列出的結果，當在模型為 Cox 比例風險模型之下的相對 KL 損失其值較小，所以根據我們所提出的模型選擇準則，在 PBC 資料中我們選擇了 Cox 比例風險模型做為配適模型。其模型如下：

$$h(t|Z) = h_0(t)\exp\,(0.14185 \times \text{serumbilirubin})$$

由配適的模型可知，當血清膽紅素每增加一單位，得到原發性膽汁性肝硬化的風險會上升至 1.15241 倍。

而在 TRACE 資料中，根據表六中的結果可看出相對 KL 損失在 Aalen 加成模型中較小，因此我們選擇 Aalen 加成模型做為配適模型。其模型如下：

$$h(t|Z) = \beta_0(t) + \beta_1(t) \times diabetes$$

其中$\beta_1(t)$之累積參數估計$\widehat{\mathrm{B}}_1(t)$的圖形則列於圖一，由圖一可看出在所有$\mathrm{t} \in \mathcal{T}$之下，其累積參數估計皆為正數，且有隨時間上升的趨勢。因此可以判斷在整體時間上，有糖尿病之病人得到心肌梗塞的風險比無糖尿病之病人高。根據 Kolmogorov-Smirnov 檢定統計量

$$\sup_{t\in[0,\varsigma]} \left|\hat{B}(t) - \hat{B}(\varsigma)\frac{t}{\varsigma}\right|$$

在虛無假設：$\beta_1(t)$與時間無關之下，其顯著性為 0.088，由此可知在信心水準為 0.1 之下，$\beta_1(t)$與時間相關。

第六章、結論及未來研究方向

本篇文章著重在於適應性模型選則準則在Cox比例風險模型與Aalen加成模型上的應用。一般在對存活資料做模型配適時，會考慮其中共變數的特性，例如：共變數是否與時間相關等，然而在配適模型中又會衍生出其他問題，如共變數是否符合比例風險假設等，所以配適存活資料時該如何選擇模型一向是我們所關心的問題。然而此適應性模型選擇準則提供了我們一個選擇模型的概念，利用相對KL損失及資料擾動的方式，幫我們從候選模型中找出相對適合的模型。在模擬結果中也能看出此方式不論擾動尺度的大小如何都有著相當高的精確度，而且在各種參數的組合下表現也不錯。

本文只將適應性模型選擇準則應用於Cox比例風險模型及Aalen加成模型上，後續若繼續探討可以考慮其他模型，如 McKeague 和 Sasieni 模型(1994)。共變數的維度也可以探討推廣至多維。

參考文獻

Aalen,O.O.(1975). Statistical inference for a family of countingprocesses. *PhD thesis, Univ. of California,Berkeley.*

Aalen,O.O.(1978a). Nonparametric estimation of partial transitionProbabilities in multiple decrement models. Ann. Statist.6,534-545.

Aalen, O.O.(1989). A linear regression model for the analysis of lifetimes.*Statist. Med*.8,907-925.

Breslow,N.E.(1975). Analysis of survival data under the proportionalhazards model. *Int. Statist. Rev.* 43,45-58.

Cox,D.R.(1972). Regression models and life tables (with discussion).*J. Roy. Statist. Soc. Ser. B* 34,187-220

Cox,D.R.(1975). Partial likelihood.*Biometrika* 62,269-276.

Chen,C.S.,Chang,Y.M.(2011). Model selection for two-sample problemswith right-censoreddata:An application of Cox model. *Journal ofStatistical Planning and Inference,*Vol.141,pp.2120-2127.

Huang, H.C., Chen, C.S.(2007). Optimal geostatistical model selection.*J.Amer. Statist.Assoc*.102,1009-1024.

Lin, D.Y. & Ying, Z.(1994). Semiparametric analysis of general additive Multiplicative hazard models for counting processes.*Ann.Statist*.23,712-1734

McKeague, I.W. &Sasieni, P.D.(1994). A partly parametric additive riskModel. *Biometrika* 81,501-514.

Shen, X., Huang, H.C., Ye, J.(2004). Adaptive model selection andassessment for exponential family models. *Technometrics* 46,306-317.

Shen, X.,Huang, H.C.(2006). Optimal model assessment, selection, andCombination. *J.Amer.Statist.Assoc*.101,554-568.

Ye, J.(1998). On measuring and correcting the effects of data mining andModel selection.*J.Amer.Statist.Assorc*.93,120-131.

附錄

表一、Cox 比例風險模型的平均 KL 損失等一覽表

真實模型：Cox 比例風險模型$\beta = 3.5$					
擾動尺度	精確度	KL loss		GDF	
		Cox	Aalen	Cox	Aalen
0.1	0.905	65.094 (0.101)	88.517 (0.619)	0.060 (8.538×10^{-4})	0.135 (7.178×10^{-3})
0.3	0.923	65.140 (0.101)	88.127 (0.610)	0.044 (4.110×10^{-4})	0.091 (2.877×10^{-3})
0.5	0.922	65.131 (0.096)	89.240 (0.618)	0.049 (4.110×10^{-4})	0.091 (2.340×10^{-3})
0.7	0.912	65.195 (0.097)	88.542 (0.611)	0.053 (3.794×10^{-4})	0.103 (2.214×10^{-3})
0.9	0.915	65.420 (0.101)	89.867 (0.633)	0.055 (3.794×10^{-4})	0.104 (2.119×10^{-3})

注：在不同的擾動尺度下，以 Cox 比例風險模型作生成的 1000 次重複的精確度，在配適 Cox 比例風險模型及 Aalen 加成模型 KL loss 及 GDF 的平均，括弧中的數字為標準差。

表二、Aalen 加成模型的平均 KL 損失等一覽表

真實模型：Aalen 加成模型					
擾動尺度	精確度	KL loss		GDF	
		Cox	Aalen	Cox	Aalen
0.1	0.872	74.439 (0.173)	73.720 (0.261)	0.022 (1.012×10^{-3})	0.051 (5.945×10^{-3})
0.3	0.881	74.439 (0.171)	73.826 (0.264)	0.023 (6.641×10^{-4})	0.050 (1.929×10^{-3})
0.5	0.876	74.604 (0.169)	73.990 (0.277)	0.034 (5.059×10^{-4})	0.055 (1.359×10^{-3})
0.7	0.867	74.369 (0.175)	73.957 (0.273)	0.044 (4.427×10^{-4})	0.066 (1.328×10^{-3})
0.9	0.867	74.373 (0.168)	73.740 (0.267)	0.053 (4.743×10^{-4})	0.073 (1.233×10^{-3})

注：在不同的擾動尺度下，以 Aalen 加成模型作生成的 1000 次重複的精確度，在配適 Cox 比例風險模型及 Aalen 加成模型 KL loss 及 GDF 的平均，括弧中的數字為標準差。

表三、Cox 比例風險模型在不同參數組合下的模型選擇

真實模型：Cox 比例風險模型$\beta = 3.5$					
參數	精確度	KL loss		GDF	
		Cox	Aalen	Cox	Aalen
$\lambda = 0.3$ $\theta = 1.2$	0.922	65.131 (0.095)	89.240 (0.618)	0.049 (4.111×10^{-4})	0.091 (2.34×10^{-3})
$\lambda = 0.3$ $\theta = 0.7$	0.927	65.158 (0.096)	87.622 (0.577)	0.040 (3.130×10^{-4})	0.079 (2.023×10^{-3})
$\lambda = 0.3$ $\theta = 2.0$	0.924	66.950 (0.101)	92.365 (0.660)	0.058 (3.415×10^{-3})	0.108 (2.561×10^{-3})
$\lambda = 0.8$ $\theta = 1.2$	0.989	64.303 (0.083)	94.686 (0.614)	0.062 (4.743×10^{-4})	0.111 (2.593×10^{-3})
$\lambda = 0.2$ $\theta = 1.2$	0.850	65.588 (0.103)	84.683 (0.593)	0.041 (3.478×10^{-4})	0.078 (2.276×10^{-3})
$\lambda = 0.8$ $\theta = 2.0$	0.992	67.364 (0.090)	72.822 (0.664)	0.027 (5.376×10^{-4})	0.056 (2.498×10^{-3})

表四、Aalen 加成模型在不同參數組合下的模型選擇

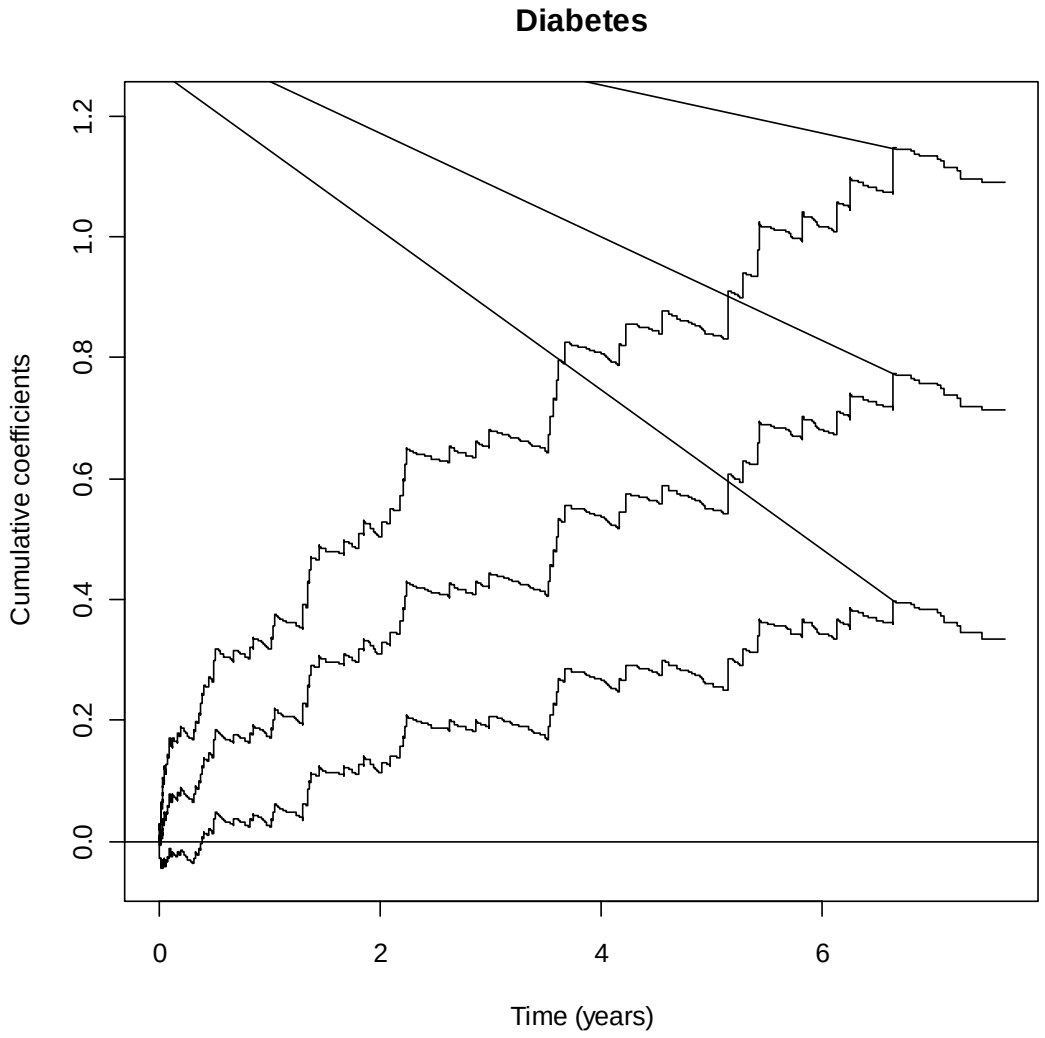
真實模型：Aalen 加成模型					
參數	精確度	KL loss		GDF	
		Cox	Aalen	Cox	Aalen
$\beta_0 = 0.1$	0.876	74.604	73.990	0.034	0.055
$\beta_1 = 1.5$		(0.172)	(0.261)	(1.011×10^{-3})	(1.359×10^{-3})
$\beta_0 = 1.2$	0.740	61.927	61.822	0.058	0.068
$\beta_1 = 1.5$		(0.099)	(0.113)	(5.05×10^{-4})	(6.957×10^{-4})
$\beta_0 = 0.3$	0.853	71.021	70.689	0.040	0.060
$\beta_1 = 1.5$		(0.148)	(0.217)	(4.743×10^{-4})	(1.201×10^{-3})
$\beta_0 = 0.1$	0.921	84.606	82.986	0.024	0.046
$\beta_1 = 0.8$		(0.258)	(0.328)	(6.324×10^{-4})	(1.391×10^{-3})
$\beta_0 = 0.1$	0.838	69.699	69.489	0.041	0.015
$\beta_1 = 2.3$		(0.135)	(0.248)	(4.743×10^{-4})	(1.296×10^{-3})

表五、原發性膽汁性肝硬化資料模型選擇結果

PBC 資料集以血清膽紅素(serum bilirubin)作為共變數之模型選擇			
待選模型	KL 損失	GDF	最後選擇
Cox 比例風險模型	262.2715	0.02408578	✓
Aalen 加成模型	266.1248	0.01453703	

表六、急性心肌梗塞資料模型選擇結果

TRACE 資料集以糖尿病(diabetes)作為共變數之模型選擇			
待選模型	KL 損失	GDF	最後選擇
Cox 比例風險模型	257.5052	0.01387769	
Aalen 加成模型	257.3500	0.01425455	✓



圖一、TRACE 資料中配適有無糖尿病共變數之 Aalen 模型參數估計及 95%信賴區間圖形