東海大學統計系碩士班

碩士論文

指導教授：黃連成 博士

不同進入時間的 Cox 迴歸模型之參數檢定

研究生：劉宜昆

中華民國九十三年七月十三日

摘　要

論文題目：不同進入時間的 Cox 迴歸模型之參數檢定

指導教授：黃連成教授

研究生：劉宜昆

論文摘要內容：

早期以 Cox 迴歸模型來分析具有不同進入時間的設限資料，作法是將進入時間調整成相同，只利用存活時間的資訊，本文對於具有不同進入時間的 Cox 迴歸模型，以 Bilias，Gu 和 Ying（1997）的定理爲基礎，提出檢定參數的檢定統計量，且電腦模擬的結果也顯示型 I 錯誤靠近假設的顯著水準，這符合假設，表示文中提出的檢定統計量合理，可以在不同的日曆時間對具有不同進入時間的設限資料檢定參數。

本文也說明了具不同進入時間模型的檢定統計量爲相同進入時間模型的推廣，並且可以用來解釋之前若進入時間是不同而將其調整成相同進入時間的資料來分析的作法是合理的。

誌 謝 辭

本論文幸蒙恩師 黃老師連成之悉心指導，在恩師的諄諄教誨下，才能順利完成，師恩浩翰，無以回報，在此謹向恩師致上最高的敬意！

本論文於審查期間，承蒙 沈教授葆聖及 紀教授美智於口試時，對學生論文諸多不足之處惠賜指正，並提供學生許多寶貴的意見，使得論文能更臻完整，特此感謝。

目 錄

第一章 緒 論

第 1.1 節 背景及研究動機

Cox（1972）迴歸模型（regression model）理論已經發展多時，討論此模型的相關文獻資料也很豐富，在設限資料（censored data）分析時被廣泛採用，包括醫學、財經、工業等領域，此模型主要分析影響因子（covariate）或稱爲解釋變數（explanatory variable）和風險（hazard）的關係。舉例來說，當運用在醫學上時，可以利用此模型來研究血壓、紅血球、體溫等因子和存活時間的關係，找出影響疾病的重要因素，真正達到預防、治療的目的。應用在財經方面時，能以公司的多項財務比率當作解釋變數，分析其和公司的實際存活時間的關係，由 Cox（1972）迴歸模型建立一套危機預警模式，當公司處於如何的狀況之下是很危險的，即公司”死亡”的機率很高，讓公司主管和投資大眾對公司的現況及遠景有實際的瞭解。Cox 迴歸模型的另一個特色是對有興趣的解釋變數在不同的水準之下，固定其他變數，比較其相對風險，即風險比（hazard ratio），與時間無關。

Cox（1975）介紹部分概似（partial likelihood）的基本觀念，並以實際的算式推導說明爲何用部分概似估計模型中的參數，包括用完整概似函數會遇到的問題及部分概似的優點。Andersen 和 Gill（1982）利用計數過程（counting process）的大樣本性質應用在 Cox 迴歸模型上，當所有觀察個體具有相同進入時間（entry time），則在時間 t 的部分概似分數（partial likelihood score）爲一個鞅（martingale），根據標準鞅中央極限定理和其他相關漸進定理推導 Cox 迴歸模型中參數估計的漸進分配及其它統計量的漸進性質。

早期Cox迴歸模型理論發展都假設所有個體的進入時間相同，然而，在各個領域中，常遇到的設限資料並非同時開始觀察，例如在醫學上，所有被觀察的病患並非同時發病；在工業上，做實驗時可能所有實驗單位不是同時開始實驗，即所有觀察個體具有不同進入時間。當然，在這種情況下將所有的觀察個體調整成同時進入實驗，並引用一般進入時間相同的理論加以分析是其中一個方法，但是，這和原本的資料不同，是經過調整的，不甚合理。對於具有不同進入時間的設限資料，早期以無母數的方法加以分析，相關理論文獻也很豐富。以Cox模型來分析具有不同進入時間的設限資料則是近幾年才開始，這種情況下在時間 t 的部分分數過程不再是鞅，所以不適用鞅的相關定理推導Cox模型的漸進性質，這使得在做相關檢定會有許多問題。對具有不同進入時間的Cox模型，考慮兩個參數的分數過程（two-parameter score process）取代原本的部分概似分數來作參數估計和檢定。Sellke和Siegmund（1983）證明在某些條件下對角過程（diagonal process）近似鞅，並弱收斂（converges weakly）到布朗運動（Brownian motion）。Gu和Lai（1991）推導出兩族群模型（two-sample model）的兩個參數分數過程的弱收斂性質。Bilias，Gu和Ying（1997）以經驗過程（empirical process）理論，來分析具有不同進入時間的Cox迴歸模型。

在這篇論文裡，主要引用 Bilias，Gu 和 Ying（1997）的定理，對於具不同進入時間的 Cox 迴歸模型在日曆時間（calendar time） t 時提出估計參數的檢定統計量，以電腦模擬的方式檢查結果是否符合假設，即檢定統計量是否合理。同時驗證具不同進入時間的模型爲相同進入時間的推廣，利用文中提出的檢定統計量亦可對具有相同進入時間的設限資料進行參數檢定。

第 1.2 節 本文架構

本文主要介紹具不同進入時間的 Cox 迴歸模型，提出一些參數的檢定統計量及其相關的模擬結果，這裡分爲四章來討論。

第一章，簡單介紹 Cox 迴歸模型的起源及相關文獻，說明本文研究背景及想要解決的問題。第二節大概敘述本文架構。

第二章的第一節介紹 Cox 的迴歸模型及 Andersen 和 Gill（1982）利用計數過程所推導統計量的漸進性質及檢定統計量。第二節介紹 Bilias，Gu 和 Ying（1997）具有不同進入時間的 Cox 迴歸模型及提出檢定模型參數的檢定統計量，並在第三節中討論這兩個模型和檢定統計量的關係。

第三章，以電腦模擬方式估計參數並驗證第二章所提出的檢定統計量是否合適。

第四章，總結這裡所討論的問題和結論，並提出後續可考慮的相關問題。

第二章 模型介紹及參數檢定

第 2.1 節 具相同進入時間的迴歸模型

對於設限資料，我們關心的是（i）失敗時間的分配（ii）存活函數（ii）風險函數或是在相同失敗時間的風險比（hazard ratio）。分析資料時可以假設失敗時間的分配服從某個參數化模型（parametric model），這樣的優點是分配形式確定，利用失敗時間的分配和存活函數及風險間的關係，只要估計模型中的參數，模型中的重要性質就能推導出，缺點是在一般情況之下，失敗時間的分配是未知的，應該做模型適合度檢定，檢查模型是否適當。另一方面，用無母數的方法估計存活函數及風險函數也很常見，相關理論和漸進定理已經發展多時，例如以 K-M 估計式（Kaplan-Meier estimate）來估計存活函數…，等等。

目前在很多應用上，使用 Cox（1972）迴歸模型來分析設限資料的頻率相當高。在這個模型中失敗時間（T）的風險（ l ）和 p 維影響因子向量（z）關係爲：

$$\exp(b^{t}\mathrm{z})\, l_0(\mathrm{t}) \qquad (2\text{-}1)$$

其中 b 爲 p 維的迴歸參數， $l_0(t)$ 爲基礎風險函數（baseline hazard function），是所有解釋變數爲零時的風險。此模型爲半參數模型（semi-parametric model），和參數有關的只有影響因子的部分，基礎風險函數是和參數無關的，這也是相乘風險模型，因爲風險爲指數的相乘形式。Cox 迴歸模型通常被稱爲風險成比例模型（proportional hazard model），當兩個解釋變數分別是 Z 和 Z^{*}的個體在失敗時間 t 的風險比爲

$$\frac{l(\mathrm{t};\mathrm{z})}{l(\mathrm{t};\mathrm{z}^{*})}=\frac{\exp\{b^{t}Z\}\,l_0(t)}{\exp\{b^{t}Z^{*}\}\,l_0(t)}=\exp\left[\sum_{k=1}^{p} b_k(z_1-z_k^{*})\right] \qquad (2\text{-}2)$$

是個常數，表示不同個體在相同失敗時間的風險是成比例的。這個特性也是此模型的優點之一，風險比的形式相當簡潔易求。例如， z_1 變數表示治療效應（ $z_1=1$ 接受治療， $z_1=0$ 接受安慰劑（placebo）），當其他變數值都相同時的風險比 $\frac{l(\mathrm{t};\mathrm{z})}{l(\mathrm{t};\mathrm{z}^*)}=\exp(b_1)$，即接受治療和安慰劑的個體在失敗時間 t 的風險比固定是 $\exp(b_1)$。

以下引用 Andersen 和 Gill（1982）以計數過程的方法介紹相關的檢定統計量，令

T_i 爲失敗時間， C_i 爲設限時間（censoring time）， $\mathrm{z}_i(t)$ 爲 $p\times1$ 影響因子向量，是時間的函數

$$a\wedge b=\min\{a,b\}$$

$$\mathring{T}_i(t)=T_i\wedge C_i\wedge t$$

$$\Delta_i(t)=I_{(T_i\le C_i\wedge t)}$$

$$N_i(t)=I_{(T_i\le C_i\wedge t)}$$

$$Y_i(t)=I_{(T_i\wedge C_i\ge t)}$$

$$R_i(t)=\{j:\mathring{T}_j(t)\ge\mathring{T}_i(t)\}$$

這裡的 $N_i(t)$ 是個計數過程，其強度過程（intensity process）爲

$$l_i(\mathrm{t})=Y_i(t)\,l_0(\mathrm{t})\exp\{b^{t}Z_{\mathrm{i}}(t)\} \qquad (2\text{-}3)$$

且根據 Cox（1972）的論文，在時間 t 的部分概似（partial likelihood）爲

$$\mathrm{L_t}(b)=\prod_{\mathrm{i}=1}^{\mathrm{n}}\left[\frac{\exp\{\,b^{t}Z_i(\mathring{T}_i(t))\,\}}{\sum_{j\in R_i(t)}\exp\{\,b^{t}Z_j(\mathring{T}_i(t))\,\}}\right]^{\Delta_i(t)} \qquad (2\text{-}4)$$

對部分概似取對數之後微分，可得到部分概似分數（partial likelihood score）

$$U(b,t)=\sum_{i=1}^{n}\int_0^t\left[Z_i(s)-\frac{\sum_{j=1}^{n}Y_j(s)Z_j(s)\mathrm{e}^{b^tZ_j(s)}}{\sum_{j=1}^{n}Y_j(s)\mathrm{e}^{b^tZ_j(s)}}\right]dN_i(s) \qquad (2\text{-}5)$$

再對部分概似分數微分取負號，可得

$$\Sigma(b,t)=\sum_{i=1}^{n}\int_0^t\left(\frac{\sum_{j=1}^{n}Y_j(s)Z_j(s)^{\otimes 2}\mathrm{e}^{b^tZ_j(s)}}{\sum_{j=1}^{n}Y_j(s)\mathrm{e}^{b^tZ_j(s)}}-\left(\frac{\sum_{j=1}^{n}Y_j(s)Z_j(s)\mathrm{e}^{b^tZ_j(s)}}{\sum_{j=1}^{n}Y_j(s)\mathrm{e}^{b^tZ_j(s)}}\right)^{\otimes 2}\right)dN_i(s)$$

（2-6）

對於向量 a， $a^{\otimes 2}$ 的意思是 aa^t 。而滿足 $U(b,t)=0$ 的解即是 b 在時間 t 的最大概似估計量，記作 $\hat{\beta}$ ，利用鞅及計數過程的方法可推導部分概似分數的漸進分配，並得到 $\hat{\beta}$ 的漸進分配，有了參數的漸進分配就可以對參數進行檢定。

以下介紹 Cox 模型中幾個檢定模型參數的統計檢定量，包括部分概似分數檢定統計量（the partial likelihood score test statistic）、Wald 檢定統計量(Wald test statistic)和部分概似比率檢定統計量(the partial likelihood ratio test statistic)。

假設 b 的維度爲 p，虛無假設 $H_0:b=b_0$ 和對立假設 $H_1:b\neq b_0$

（一）部分概似分數檢定統計量

由 Andersen 和 Gill（1982）定理 3.2 的證明中可得到，在虛無假設下

$$n^{-1/2}U(b_0,\mathrm{t})\rightarrow_D N(0，\Psi) \qquad (2\text{-}7)$$

因 $n^{-1}\Sigma(b_0,\mathrm{t})$ 和 $n^{-1}\Sigma(\hat{\beta},\mathrm{t})$ 是 Ψ 的一致性（consistent）估計量，故可得

$$\begin{aligned}\mathrm{S}&=\mathrm{n}^{-1/2}U^t(b_0,t)\left(n\Sigma^{-1}(b_0,t)\right)\mathrm{n}^{-1/2}U(b_0,t)\\&=U^t(b_0,t)\Sigma^{-1}(b_0,t)U(b_0,t)\\&\approx c^2(p)\end{aligned} \qquad (2\text{-}8)$$

（二）Wald 檢定統計量

由 Andersen 和 Gill（1982）定理 3.2 可得知在虛無假設成立時

$$n^{1/2}(\hat{\beta} - b_0) \rightarrow_D N(0, \Psi^{-1}) \qquad (2\text{-}9)$$

所以

$$\begin{aligned} W &= n^{1/2}\left(\hat{\beta} - b_0\right)^t \left(n^{-1}\Sigma(\hat{\beta}, t)\right) n^{1/2}\left(\hat{\beta} - b_0\right) \\ &= \left(\hat{\beta} - b_0\right)^t \Sigma(\hat{\beta}, t)\left(\hat{\beta} - b_0\right) \\ &\approx c^2(p) \end{aligned} \qquad (2\text{-}10)$$

（三）部分概似比率檢定統計量

$$\begin{aligned} L &= 2\{\ln L_t(\hat{\beta}) - \ln L_t(b_0)\} \\ &\approx c^2(p) \end{aligned} \qquad (2\text{-}11)$$

以上爲具有相同進入時間 Cox 迴歸模型檢定參數常用的檢定統計量。

第 2.2 節 具不同進入時間的迴歸模型

Cox 迴歸模型早期發展時考慮所有觀察個體同時進入實驗，但是在很多實際情況下，觀察個體可能具有不同進入時間，此時之前模型便不適用，這一章引用 Bilias，Gu 和 Ying（1997）具有不同進入時間的模型，文中以兩個參數分數過程（two-parameter score process）來估計參數，並推導出相關收斂定理來檢定參數。

令t_i（≥ 0）爲第 i 個個體的進入時間，i 從 1 到 n。假設（t_i，T_i，C_i，$Z_i^{'}$）爲獨立隨機向量，在給定t_i、C_i和$Z_i(u)$（$u \leq s$）之下，T_i的條件風險率（conditional hazard rate）爲

$$\exp\{b^t Z_i(s)\} l_0(s) \qquad (2\text{-}12)$$

l_0(s)爲基礎風險函數。由此可知若給定t_i、C_i和$Z_i(u)$，條件風險率只和影響因子及基礎風險函數有關。令

$a \wedge b = \min\{a,b\}$ ，$a^+ = \max\{0,a\}$

$\Delta_i(t) = I_{(T_i \leq C_i \wedge (t-t_i)^+)}$

$\mathring{T}_i(t) = T_i \wedge C_i \wedge (t-t_i)^+$

$R_i(t) = \{j : 1 \leq j \leq n, \mathring{T}_j(t) \geq \mathring{T}_i(t)\}$

根據 Cox 風險成比例模型，在日程時間 t 的部分概似爲

$$L_t(b) = \prod_{i=1}^{n} \left[\frac{\exp\{ b^t Z_i(\mathring{T}_i(t)) \}}{\sum_{j \in R_i(t)} \exp\{ b^t Z_j(\mathring{T}_i(t)) \}} \right]^{\Delta_i(t)} \qquad (2\text{-}13)$$

和一般 Cox 迴歸模型中計數過程表示方法不同的是在此必須考慮兩種不同時間的類型，日程時間 t 和存活時間 s，在正常情況下 $t \geq s$。

令

$$N_i(t,s)=I_{(T_i\le C_i\wedge(t-t_i)^+\wedge s)}\ ,\ Y_i(t,s)=I_{(T_i\wedge C_i\wedge(t-t_i)^+\ge s)}$$

定義兩個參數的分數過程（two-parameter score process）

$$U(b;t,s)=\sum_{i=1}^{n}\int_0^s\left[Z_i(u)-\overline{Z}(b;t,u)\right]N_i(t,du) \qquad (2\text{-}14)$$

其中

$$\overline{Z}(b;t,s)=\frac{\sum_{l=1}^{n}Z_l(s)\exp\{\ b^tZ_l(s)\ \}Y_l(t,s)}{\sum_{l=1}^{n}\exp\{\ b^tZ_l(s)\ \}Y_l(t,s)} \qquad (2\text{-}15)$$

同理，令$U(b;t,s)=0$可得到 b 在日曆時間 t 和存活時間 s 的最大概似估計值，記作 $\hat{b}$(t,s)。觀察個體具有不同進入時間的分數過程不是個鞅，故不適用鞅的相關定理，此時以經驗過程理論推導分數過程的漸進分配。以下介紹具不同進入時間 Cox 迴歸模型的檢定統計量。

假設 b 的維度爲 p，虛無假設 $H_0:b=b_0$和對立假設 $H_1:b\neq b_0$，由 Bilias，Gu 和 Ying（1997）定理 2.2 證明兩個參數的分數過程分佈收斂到高斯隨機體（Gaussin random field）和定理 4.2 證明 $\hat{b}(t,s)$也是分佈收斂到高斯隨機體，這裡根據這兩個定理提出以下檢定統計量，定理詳細敘述請參閱附錄。

（一）部分概似分數檢定統計量

$$\begin{aligned}\text{S}&=\frac{1}{\sqrt{\text{n}}}U^t(b_0;t,s)n\Sigma^{-1}(b_0;t,s)\frac{1}{\sqrt{\text{n}}}U(b_0;t,s)\\&=U^t(b_0;t,s)\Sigma^{-1}(b_0;t,s)U(b_0;t,s)\end{aligned} \qquad (2\text{-}15)$$

因爲 Bilias，Gu 和 Ying（1997）定理 2.2 $U(b_0;t,s)$收斂到高斯隨機體中的共變異函數包含未知的基礎風險函數且形式太複雜，這裡以$n^{-1}\Sigma(b_0;\text{t,s})$估計共變異函數。

其中

$$\Sigma(b;t,s)$$
$$=\sum_{i=1}^{n}\int_0^s\left[\frac{\sum_{l=1}^{n}Z_l^{\otimes 2}(u)Y_l(t,u)\exp\{\ b^tZ_l(u)\ \}}{\sum_{l=1}^{n}Y_l(t,u)\exp\{\ b^tZ_l(u)\ \}}-\left(\frac{\sum_{l=1}^{n}Z_l(u)Y_l(t,u)\exp\{\ b^tZ_l(u)\ \}}{\sum_{l=1}^{n}Y_l(t,u)\exp\{\ b^tZ_l(u)\ \}}\right)^{\otimes 2}\right]N_i(t,du)$$

（2-16）

（二） Wald 檢定統計量

$$\begin{aligned}W&=\sqrt{n}\{\ \hat{\beta}(t,s)-b_0\ \}^t n^{-1}\Sigma(\hat{\beta};t,s)\sqrt{n}\{\ \hat{\beta}(t,s)-b_0\ \}\\&=\{\ \hat{\beta}(t,s)-b_0\ \}^t\Sigma(\hat{\beta};t,s)\{\ \hat{\beta}(t,s)-b_0\ \}\end{aligned}\qquad（2\text{-}17）$$

這裡以 $n\Sigma^{-1}(\hat{\beta};t,s)$ 作爲 $\hat{\beta}$ 收斂到高斯隨機體中共變異函數的估計。

第 2.3 節 模型及檢定的關係

具相同進入時間和不同進入時間的 Cox 迴歸模型部分概似表示式相同，都是

$$L_t(b)=\prod_{i=1}^{n}\left[\frac{\exp\{\, b^t Z_i(\mathring{T}_i(t))\,\}}{\sum_{j\in R_i(t)}\exp\{\, b^t Z_j(\mathring{T}_i(t))\,\}}\right]^{\Delta_i(t)}$$

只是兩個模型中的$\mathring{T}_i$和$\Delta_i(t)$的定義方式不同，但若把後者模型的進入時間t_i調整成 0 時，則兩者皆一樣，即考慮進入時間模型的$\mathring{T}_i(t)=T_i\wedge C_i\wedge(t-t_i)^+$在 t_i都爲 0 情況下簡化爲進入時間相同模型的$\mathring{T}_i(t)=T_i\wedge C_i\wedge t$，$\Delta_i(t)=I_{(T_i\le C_i\wedge(t-t_i)^+)}$簡化爲$I_{(T_i\le C_i\wedge t)}$，表示所有觀察個體同時進入，兩個模型的部分概似就完全相同了。同理，這兩個模型的檢定統計量結構亦大致相同，若將t_i都視爲零時，兩個模型的部分概似分數檢定統計量和 Wald 檢定統計量也會相同。

由此說明具有不同進入時間模型的部分概似分數檢定統計量和 Wald 檢定統計量是相同進入時間模型的推廣，本文也以模擬的方式來驗證。

在具有不同進入時間的 Cox 模型還沒被提出之前，對於具有不同進入時間的設限資料如果以 Cox 模型來分析，作法是將進入時間調整成相同，或是忽略，只以存活時間來分析，在此也說明了這樣的作法的合理性。

第三章 電腦模擬

對於具有相同進入時間的 Cox 模型，已經有許多知名的電腦軟體，如：SAS、S-PLUS，提供現成的指令可以對參數進行估計和檢定，這使得利用 Cox 模型來分析設限資料變得簡單容易，這也是 Cox 模型被頻繁採用的原因之一，另一方面，由於具有不同進入時間的 Cox 模型發展時間並不長，模型中若考慮進入時間而想要估計參數，目前還不是相當方便，所以本章以模擬的方式對於具有不同進入時間的設限資料 Cox 迴歸模型在不同的日曆時間 t 進行參數的估計和檢定。在此我們考慮 p＝2 的情形，即迴歸模型中有兩個參數：(b_1, b_2)。

本文以下列步驟進行模擬

步驟 1：每組資料樣本數設定 n=100，假設 t_i 、 C_i 和 Z_1、Z_2 服從某些特定分配（見下表），以給定的分配來產生 100 個資料，作爲每個個體的進入時間、設限時間和解釋變數，這裡假設 $l_0(\mathrm{t})=1$ 和兩組參數 (b_1, b_2)爲（0,0）和（3,0.5），在 t_i 、 C_i 、 Z_1、Z_2 和參數給定後，第 i 個個體的條件風險率便確定了，爲 $\exp(b^t \mathrm{z}_i)$。這裡我們以 t_i 和 Z_1、Z_2 不同分配的組合來產生資料。

步驟 2：當條件風險率爲一常數 a 時，可得到失敗時間 T 的機率密度函數爲 ae^{-at}，是參數 a 的指數分佈，所以第 i 個個體失敗時間爲指數分配 $\mathrm{Exp}\left(\exp(b^t \mathrm{z}_i)\right)$，由此分配產生第 i 個個體的失敗時間。

步驟 3：每個個體的進入時間、設限時間、失敗時間和解釋變數都產生之後，在日程時間 t 令（2-13）式兩個參數分數過程 $U(b;t,s)=0$，在此考慮 s=t 的情況，滿足的解即是此時 b 的最大概似估計值，將以上步驟重複 1000 次，計算時間 t 參數估計值的期望值當作估計值

並計算標準差。

步驟 4：檢定部分，我們兩個虛無假設爲 H_0: $b_1 = 0, b_2 = 0$ 和 H_0: $b_1 = 3, b_2 = 0.5$，在給定顯著水準 0.05 之下，將每一組產生的資料和估計出的參數帶入檢定統計量 S（2-15）、W（2-17），計算檢定統計量是否落於棄卻域，判斷是否拒絕虛無假設。

步驟 5：將步驟 4 重複 1000 次，計算拒絕 H_0 的比例，若比例接近 0.05 表示檢定和假設相符，即檢定統計量是合理的。

步驟一假設 t_i、C_i 和 Z_i 服從下列分配：

<table>
<tr><td>t_i（進入時間）</td><td colspan="2">u 0（即所有個體進入時間相同）
u $U(0,1)$</td></tr>
<tr><td rowspan="2">Z_1、Z_2（解釋變數）</td><td>離散</td><td>連續</td></tr>
<tr><td>u $P(z_i) = \begin{cases} 0.6, z_i = 1 \\ 0.4, z_i = 2 \end{cases}$
u $P(z_i) = \begin{cases} 0.3, z_i = 1 \\ 0.2, z_i = 2 \\ 0.5, z_i = 3 \end{cases}$</td><td>u $U(0,50)$
u N(0.1,0.05)</td></tr>
<tr><td>C_i（設限時間）</td><td colspan="2">u 50
u N(1.2 , 0.1)
u N(5 , 0.5)
u N(4 , 0.5)</td></tr>
</table>

具有不同進入時間的 Cox 迴歸模型之參數估計和檢定模擬結果

表一

假設參數 b_1=3 , b_2=0.5

$t_i \sim U(0,1)$

	t	0.4	0.6	0.9	1.5
$P(z_1)=\begin{cases}0.6, z_1=1\\0.4, z_1=2\end{cases}$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	4.728678 （3.342026）	3.159356 （0.594736）	3.085749 （0.409664）	3.071418 （0.384826）
	β_2	0.555027 （0.654679）	0.511018 （0.173655）	0.509346 （0.138103）	0.510663 （0.129160）
	W (%)	0.033000	0.032000	0.044000	0.051000
	S (%)	0.066000	0.041000	0.045000	0.053000
$Z_1 \sim N(0.1,0.05)$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	3.304054 （5.390606）	3.200263 （3.654875）	3.075188 （2.715386）	3.035654 （2.238447）
	β_2	0.526358 （0.340616）	0.511112 （0.225497）	0.510914 （0.167476）	0.506598 （0.136596）
	W (%)	0.037000	0.038000	0.046000	0.048000
	S (%)	0.063000	0.060000	0.052000	0.052000
$Z_1 \sim U(0,50)$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	3.761940 （1.801270）	3.248345 （0.788301）	3.108394 （0.454026）	3.087915 （0.397024)
	β_2	0.651107 （0.772742）	0.536612 （0.34037）	0.514915 （0.220551）	0.511566 （0.198640)
	W (%)	0.052000	0.051000	0.049000	0.046000
	S (%)	0.060000	0.055000	0.053000	0.054000
$Z_1 \sim N(0.1,0.05)$ $Z_2 \sim U(0,50)$ $C_i \sim N(1.2,0.1)$	β_1	2.877059 （4.991604）	2.906635 （3.522668）	2.973242 （2.600017）	2.969684 （2.385790）
	β_2	0.527686 （0.104488）	0.513727 （0.067819）	0.507316 （0.048738）	0.506241 （0.046232）
	W (%)	0.049000	0.047000	0.061000	0.057000
	S (%)	0.062000	0.065000	0.072000	0.059000

t ：日曆時間

表中數值爲模擬 1000 次之參數估計平均值，括號內爲標準差。

W (%)：模擬 1000 次中 Wald 檢定拒絕 $H_0: b_1=3, b_2=0.5$ 的百分比。

S (%)：模擬 1000 次中部分概似分數檢定拒絕 $H_0: b_1=3, b_2=0.5$ 的百分比。

表二

假設參數 $b_1=b_2=0$

$t_i \sim U(0,1)$

	t	1	2	5	30
$P(z_1)=\begin{cases}0.6, z_1=1\\0.4, z_1=2\end{cases}$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	-0.003752	0.004272	0.003857	0.004437
		(0.350264)	(0.243329)	(0.216426)	(0.216279)
	β_2	-0.000815	-0.004531	-0.003602	-0.003673
		(0.198774)	(0.134819)	(0.119120)	(0.118933)
	W (%)	0.025000	0.036000	0.043000	0.043000
	S (%)	0.056000	0.055000	0.055000	0.057000
$Z_1 \sim N(0.1,0.05)$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	0.013112	-0.020596	-0.058580	-0.051850
		(3.411691)	(2.439252)	(2.171855)	(2.164812)
	β_2	-0.005892	-0.005607	-0.007220	-0.007225
		(0.198065)	(0.141851)	(0.124625)	(0.124024)
	W (%)	0.022000	0.050000	0.059000	0.059000
	S (%)	0.065000	0.049000	0.055000	0.053000
$Z_1 \sim U(0,50)$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	-0.000678	-0.000221	-0.000296	-0.000194
		(0.012466)	(0.008498)	(0.007989)	(0.007565)
	β_2	0.005357	-0.000676	-0.003280	-0.004804
		(0.198660)	(0.138298)	(0.130665)	(0.123290)
	W (%)	0.021000	0.046000	0.047000	0.055000
	S (%)	0.061000	0.047000	0.055000	0.056000
$Z_1 \sim N(0.1,0.05)$ $Z_2 \sim U(0,50)$ $C_i \sim N(5,0.5)$	β_1	-0.077962	-0.088456	-0.074660	-0.074938
		(3.626169)	(2.475420)	(2.226690)	(2.223394)
	β_2	-0.000139	-0.000158	-0.000369	-0.000368
		(0.012060)	(0.008122)	(0.007430)	(0.007423)
	W (%)	0.020000	0.044000	0.054000	0.054000
	S (%)	0.064000	0.051000	0.057000	0.054000

W (%)：模擬 1000 次中 Wald 檢定拒絕 $H_0: b_1=b_2=0$ 的百分比。

S (%)：模擬 1000 次中部分概似分數檢定拒絕 $H_0: b_1=b_2=0$ 的百分比。

表三

假設參數 $b_1=b_2=0$

$t_i=0$

	t	1	2	2.5	30
$P(z_1)=\begin{cases}0.6, z_1=1\\0.4, z_1=2\end{cases}$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	-0.006836 （0.255080）	0.000044 （0.219742）	0.000150 （0.214796）	0.000475 （0.212005）
	β_2	-0.000367 （0.147626）	-0.001135 （0.126965）	-0.001879 （0.123623）	-0.001938 （0.121153）
	W (%)	0.025000	0.032000	0.037000	0.044000
	S (%)	0.062000	0.071000	0.069000	0.044000
$Z_1\sim N(0.1,0.05)$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	-0.048266 （2.689799）	-0.087819 （2.348496）	-0.091222 （2.261987）	-0.091600 （2.227624）
	β_2	0.000025 （0.145916）	0.000388 （0.126047）	-0.000969 （0.123653）	-0.002318 （0.122941）
	W (%)	0.036000	0.047000	0.050000	0.057000
	S (%)	0.060000	0.050000	0.048000	0.057000
$Z_1\sim U(0,50)$ $P(z_2)=\begin{cases}0.3, z_2=1\\0.2, z_2=2\\0.5, z_2=3\end{cases}$ $C_i=50$	β_1	0.000160 （0.009009）	0.000211 （0.007757）	0.000203 （0.007574）	0.000249 （0.007473）
	β_2	-0.000776 （0.146833）	-0.001804 （0.128103）	-0.002246 （0.125204）	-0.003284 （0.123142）
	W (%)	0.031000	0.033000	0.041000	0.047000
	S (%)	0.060000	0.057000	0.049000	0.049000
$Z_1\sim N(0.1,0.05)$ $Z_2\sim U(0,50)$ $C_i\sim N(4,0.5)$	β_1	-0.044170 （2.696009）	-0.082997 （2.355191）	-0.086570 （2.264421）	-0.083586 （2.230918）
	β_2	-0.000040 （0.008974）	0.000026 （0.007673）	-0.000030 （0.007539）	-0.000054 （0.007470）
	W (%)	0.038000	0.048000	0.050000	0.052000
	S (%)	0.053000	0.057000	0.062000	0.055000

W (%)：模擬 1000 次中 Wald 檢定拒絕 $H_0: b_1=b_2=0$ 的百分比。

S (%)：模擬 1000 次中部分概似分數檢定拒絕 $H_0: b_1=b_2=0$ 的百分比。

我們藉由電腦模擬最主要是要驗證本文提出的檢定統計量是否正確。我們的目標是在日曆時間 t 適當時，Wald 檢定和部分概似分數檢定拒絕虛無假設的比例接近我們所假設的顯著水準 0.05。

由模擬的結果可知，當日曆時間夠大時兩個檢定拒絕虛無假設的比例均能控制在 0.05 正、負 0.01 之內，說明了以這兩個檢定統計量來檢定參數是正確的，和事實相符。在具有不同進入時間的 Cox 迴歸模型，可以利用這兩個檢定統計量來檢定參數。

當日曆時間不夠大時，樣本失敗的個數較少，由於 Cox 迴歸模型中部分概似函數的特性，樣本所提供的資訊不足，導致參數估計的偏差和變異都較大，檢定的結果也不理想，即拒絕虛無假設的比例和顯著水準 0.05 相差較大，此乃 t 太小或者樣本數不夠大的原因，並非檢定統計量錯誤，至於多大的日曆時間才夠大足以正確的估計和檢定參數，這和參數、解釋變數、進入時間、設限時間都有關係，並沒有一致的答案。

第四章 結論

對於具有不同進入時間的設限資料，以 Cox 迴歸模型來分析是近幾年被提出。本文主要提出兩個檢定統計量，在不同的日曆時間檢定參數，由電腦模擬結果顯示這兩個檢定拒絕虛無假設的比例和本文設定的顯著水準 0.05 接近，符合我們的假設，驗證了這兩個檢定統計量是合理的。

早期以 Cox 模型來分析具有不同進入時間的設限資料，作法是將進入時間調整成相同來分析，本文提出的檢定統計量可以用來解釋這樣的作法。

由於 Cox 部分概似只採用失敗個體的資訊，所以當在日曆時間很小且失敗的個體太少，即大部分的個體設限或者尚未進入實驗時，樣本所提供的資訊不足，此時由部分概似分數得到的參數估計值變異很大，檢定的結果也不佳，因此雖然具有不同進入時間的 Cox 迴歸模型可以分析非同時開始觀察的設限資料，但如果觀察的時間太早或樣本數不夠，則所估計出的參數是較差的，應該將觀察時間延後，如此可以得到較好的參數估計值。

對於 Cox 模型中參數檢定的另一個常用的檢定統計量，部分概似比率檢定統計量（the partial likelihood ratio test statistic）

$$L=2\{ \ln L_t(\hat{\beta}) - \ln L_t(b_0) \} \qquad (5\text{-}1)$$

在實驗個體具有相同進入時間的情況下，L 的漸進分配爲具有自由度 p 的卡方分配，但是本文並沒有將它推廣在此模型上來檢定參數，我們相信部分概似比檢定同樣可以應用在具有不同進入時間的 Cox 迴歸模型上，如此一來，就可以和文中所提到的兩個檢定在不同情況下比較檢定力，這是日後可以探討的問題。

參考文獻

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. Ann. Statist. 10, 1100-1120.

Bilias, Y., Gu, M. and Ying, Z. (1997). Towards a general asymptotictheory for Cox model with staggered entry. Ann. Statist. 25, 662-682.

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

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

Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. Ann. Statist. 19 1403-1433.

Johansen, S. (1983). An extension of Cox's regression model. Int. Statist. Rev. 51 (to appear).

Kalbfleisch, J. D. and Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data. Wiley , New York.

Klein, J. P. and Moeschberger, M. L. (1997). Survival Analysis. Techniques for Censored and Truncated Data. Springer, New York.

Kleinbaum, D. G. (1996). Survival Analysis. A Self-Learning Text. Springer, New York.

Rebolledo, R. (1978). Sur les applications de la théorie des martingales a l'étude statistique d'une famille de processus ponctuels. Springer Lect. Notes in Mathematics 636 27-70.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. Biometrika 70 315-326.

附　錄

以下介紹 Bilias，Gu 和 Ying（1997）的論文裡，對於具有不同進入時間的 Cox 迴歸模型中分數過程及參數估計量的漸進分配定理，我們引用此定理提出具有不同進入時間 Cox 迴歸模型的部分概似分數檢定統計量和 Wald 檢定統計量來檢定參數。

令

$$\Gamma_k(t,s)=\lim_{x\to\infty}\frac{1}{n}\sum_{i=1}^{n}E[Z_i^{\otimes k}(s)Y_i(t,s)\exp\{\ b_0^tZ_i(s)\ \}]\ \ \forall t,s\in D_*\quad ，\mathrm{k}=0、1、2$$

對於向量 a， $a^{\otimes 0}=1$, $a^{\otimes 1}=a$, $a^{\otimes 2}=aa^t$ 。

$$m(t,s)=\int_0^s\left[\Gamma_2(t,u)-\frac{\Gamma_1^{\otimes 2}(t,u)}{\Gamma_0(t,u)}\right]l_0(u)du$$

1. 兩個參數分數過程的收斂性（Convergence of the two-parameter score process）

   ı　定理 2.2

   滿足正規條件下，$\frac{U}{\sqrt{n}}$ 分配收斂到 x 在 $D_*=\{(\mathrm{t}，\mathrm{s})：0\le \mathrm{s}\le \mathrm{t}\le \mathrm{t}^*\}$， x 爲期望值是零向量且共變異函數是

$$\mathrm{E}\left[x(\mathrm{t}_1,\mathrm{s}_1)x^t(\mathrm{t}_2,\mathrm{s}_2)\right]$$
$$=\int_0^{s_1\wedge s_2}\left[\Gamma_2(\mathrm{t}_1\wedge \mathrm{t}_2,u)-\frac{\Gamma_1^{\otimes 2}(\mathrm{t}_1\wedge \mathrm{t}_2,u)}{\Gamma_0(\mathrm{t}_1\wedge \mathrm{t}_2,u)}\right]l_0(u)du$$

的高斯隨機體。

2. 最大部分概似估計量的收斂性質（Convergence of the maximum partial likehood estimator）

   ı　定理 4.2

   假設 β 爲最大部分概似估計量， b_0 爲虛無假設下的參數，在滿足正規

條件之下， $\sqrt{n}\,(\hat{b}(\mathrm{t,s})-b_0)$ 分配收斂到高斯隨機體 h ，h 的期望值爲零向量且共變異函數爲

$$\mathrm{E}\{h(\mathrm{t}_1,\mathrm{s}_1)h^t(\mathrm{t}_2,\mathrm{s}_2)\}=m^{-1}(\mathrm{t}_1,\mathrm{s}_1)m(\mathrm{t}_1\wedge\mathrm{t}_2,\mathrm{s}_1\wedge\mathrm{s}_2)m^{-1}(\mathrm{t}_2,\mathrm{s}_2)$$

在本文裡，我們以 $n^{-1}\Sigma(b_0;\mathrm{t,s})$ 和 $n^{-1}\Sigma(\hat{\boldsymbol{\beta}};\mathrm{t,s})$ 作爲 $m(t,s)$ 的估計。