

漫談多維度存活分析統計方法之發展

交通大學統計所 王維菁

存活分析簡介

存活分析(survival analysis)是一門研究存活變數的統計方法。廣義的「存活變數」,可定義為由自定的時間起點到發生某個隨機事件(event)所需時間的長度,此學科因此亦被稱為「事件史分析」(event history analysis)。存活分析常被應用在生物醫學方面的研究,例如將病人腫瘤切除手術的時間訂為起點,該病人死亡(或復發)的時間訂為終點,此段時間的長短可用來衡量手術對癌症治療的效果。工業統計中的「可靠度分析」,唯一門探討產品壽命週期的統計方法(見 Meeker [1])。勞動經濟學者也常用存活分析的方法研究勞工問題,例如西元 2000 年諾貝爾經濟學獎得主 James Heckman 即曾經發表許多與 survival analysis 有關的文章。他獲獎的理由之一是以統計的方法探討勞工問題,對有關勞工的教育、社會保險、法律…等政策擬定,具有重要的影響。詳細資料請見芝加哥大學經濟系的網頁(<http://economics.uchicago.edu>)。

在蒐集存活資料時,有時無法確切的記錄所感興趣事件所發生的時間,此現象稱之為設限(censoring)。造成設限的原因可能是在進行資料分析時所關切的事件尚未發生;也可能因受測者中途離開研究(drop out)。以常見的右設限(right censoring)為例,令 T 為感興趣的存活變數, C 為設限變數,在右設線下所觀測到的變數是 $X = \min(T, C)$, $d = I(T \leq C)$ 。概括而言,存活分析提供了處理不完整設限資料的統計方法。

傳統存活分析的統計方法大致可分三個方向。第一是母數分析,即假設存活變數的母數分配模式已知。此方向較常被研究工程可靠度分析的學者採用,因為產品的物理性質可用來做模式選擇的依據。一般生物統計方面的應用,多採無母數(nonparametric)或是半母數(semi-parametric)的方式。最重要的無母數推論方法,是 Kaplan 與 Meier [2] 所提出的存活函數估計量。最著名的半母數模式,是 Cox [3] 提出的 proportional hazard model。

存活分析的研究方法在 1980 年前後產生了重大的變革,起於 Aalen [4] 將鞅理論(martingale theory)用在分析事件史資料。其主要想法如下,前述之右設限資料可改以 $I(X \leq t, d = 1)$ 計數過程(counting process)描述之。根據著名的 Doob-Meyer decomposition,可證明

$$M(t) = I(X \leq t, d = 1) - \int_0^t I(X \geq u) l(u) du$$

是個 martingale (鞅),其中 $l(u)$ 是 T 的風險函數(hazard function)。存活分析中原本繁複的大樣本性質的推導工作因鞅理論的應用而得以大為簡化(見 Fleming and Harrington [5]),同時也吸引一些原本從事機率方面研究的學者投入此較具應用色彩的研究領域。此時也有人開始將一維的問題,試著拓展到二維的平面甚至更高的維度。

多維度存活分析 — 關聯性的探討

多維度存活分析(multivariate survival analysis)主要探討數個存活變數間的聯合行為。令 $(T_1, \dots, T_k)$ 代表所感興趣的存活變數,許多研究的重點在了解這些變數之間的相關性。例如遺傳學家可以藉著研究雙胞胎壽命間的關聯性以探討基因對壽命的影響。其它的例子如研究人體成對的器官(如眼、耳、腎臟、肺、乳房)發生某個疾病時間的相關性;對具有多階段病程(multi-stage)的疾病(如愛滋病),探討每個病程所經歷時間的長度的是否有所關聯?

如何「量化」存活變數間的關聯性(即找到合適的參數、函數或模式),使其背後的機制能夠清楚的呈現,是個值得探討的課題。以二維的例子而言,一些介於-1到1的相關係數(如 Pearson correlation, kendall's tau 與 Spearman's rho 等)提供了最概括的指標。除此之外科學家往往希望更進一步研究多維存活變數在不同時間點關聯性的變化,藉此了解一些更深入的科學問題。再以雙胞胎壽命(表示為 T_1 與 T_2)為例,兩者的相關性的強弱變化隱含了基因如何影響壽命的訊息。例如若懷疑早夭的嬰兒是因為先天基因不健全,我們預期看到的現象會

是當 T_1, T_2 的值很小的時候，變數間有較強的關聯性。若猜測長壽的原因受到相當程度遺傳的影響， (T_1, T_2) 間的相關性在其值很大的時候亦應較強。Anderson, Louis, Holm and Harvald [6] 介紹一些有用的 local dependence measures。變數間的聯合行為亦可以透過機率模型描述之。因為存活變數多是非對稱的，純母數的多維分配模式 (parametric multivariate models) 一般說來應用性並不大。以二維為例，bivariate frailty model (Oakes [7]) 頗受研究者的青睞。此模式的建構是基於以下兩個假設：(1) T_1 與 T_2 的關聯性由一個看不到的 frailty 變數 (以 W 表示) 所完全解釋。因此若 W 已知， T_1 與 T_2 彼此獨立。(2) W 對 T_j ($j=1,2$) 的影響可以用 proportional hazard 的模式描述之。Frailty 變數可視為一種 random effect，以雙胞胎的例子而言，此 frailty 變數似乎代表了雙生子共有之基因或環境因素中影響死亡的因子。在假設 (1) 與 (2) 皆成立下， (T_1, T_2) 的聯合分配可表示為：

$$pr(T_1 > t_1, T_2 > t_2) = f_a \{ f_a^{-1}[S_1(t_1)] + f_a^{-1}[S_2(t_2)] \},$$

其中 $S_j(t_j) = \Pr(T_j > t_j)$ ($j=1,2$)， $f_a(\cdot)$ 是 W 的 Laplace transformation。Bivariate frailty 模式可以視為以下更廣義 copula model 的特例，後者定義如下：

$$pr(T_1 > t_1, T_2 > t_2) = C_a \{ S_1(t_1), S_2(t_2) \}.$$

文獻中許多有用的二維存活函數模式均可以表示為 copula 型態(見 Genest and Rivest[8])，例如 Clayton [9] 所提出的模式是當 W 具有 Gamma 分配時。Copula 模式最大的優點是可以將邊際分配與關聯性分開討論。描述關聯性的是 $C_a(\cdot)$ 函數，參數 a 與相關係數 Kendall's tau 常有一對一的對應關係， T_j 的邊際行為由 $S_j(\cdot)$ ($j=1,2$) 描述，通常不做母數分配的假設，因此 copula model 被認為是一種半母數模式。

多維度存活分析 — 設限下的統計推論

多維度存活資料的設限情形亦可能是多維度，而且設限的機制較一維的情形更為複雜。令 $(C_1, \dots, C_k)$ 為不感興趣的設限變數，在右設限的情形下所觀察到的資料是 $(C_1, \dots, C_k)$ 與 $(d_1, \dots, d_k)$ ，其中 $X_i = \min(T_i, C_i)$ ， $d_i = I(T_i \leq C_i)$ ($i=1, \dots, k$)。在設限的情形下欲研究 $(T_1, \dots, T_k)$ ，必須做些假設才能使問題 identifiable。一般常用的假設是令所有的 T_j 和 C_i ($i, j=1, \dots, k$) 均獨立，然

而此假設在某些應用上不盡合理(見結論)。

目前文獻上的討論仍以二維為主，以下簡要的敘述一些重要的統計方法的發展。80 年代初期統計學家嚐試將 Kaplan-Meier [2] 提出的無母數估計量推廣，用以估計二維存活函數，然而卻遭遇了始料未及的困難。後來著名的學者 Gill [10] 對此問題以 'notoriously tough' 形容之，原因如下：在二維平面或多維度空間的 '時間路徑' 並非唯一，因此很難直接仿照 Kaplan-Meier 估計量所使用的 product limit 的觀念。嚐試以 nonparametric MLE 所解得之 $pr(T_1 > t_1, T_2 > t_2)$ 估計量，連最基本的大樣本一致性 (consistency) 都不具備。文獻上最常被引用的二維存活函數估計量是 Dabrowska [11] 和 Prentice and Cai [12] 所提出的，兩者均可以延伸到更高的維度，均具有一致性，在大樣本下以 $n^{1/2}$ 的速率收斂至常態分配。然而兩者(包括一些類似的無母數估計量)皆可能估計出負的機率(見 Pruitt[13])，且計算非常繁複。美妙的鞅理論並不適用於解決多維度的問題，因為在平面上(或更高維度)並不存在絕對的順序關係 (order relationship)，因此無法建構所謂 nested filtration。此外對其它相關參數(如 Kendall's tau) 或相關函數(如 local odds ratio)的推論，無母數的推論方式往往牽涉技巧上的難度。至於有關檢定 T_j 與 T_i 間獨立性的問題則較為單純，因為在虛無假設下，檢定量的性質只與邊際分配有關(見 Hsu and Prentice [14])。

90 年代後一些統計學家嚐試在較廣義的模式假設下探討存活變數的相關性。例如以半母數 (semi-parametric) 的方法，估計 copula model 中的相關參數 a 。Shih and Louis [15] 所提出的做法是先把 $S_j(T_j)$ ($j=1,2$) 視為兩個彼此相關但具 Uniform 邊際分配的隨機變數， $C_a(\cdot)$ 視為兩者的聯合存活函數，以此建構概似函數。然而因 $S_j(T_j)$ ($j=1,2$) 無法直接觀測到且可能受到設限，因此作者建議以 $S_j(T_j)$ 的 Kaplan-Meier 估計量取代之，再對所謂 '擬概似函數' (pseudo-likelihood function) 求取極大值。在實證研究時，研究者往往面臨 $C_a(\cdot)$ 該如何選取的問題，對此問題 Wang and Wells [16] 提出模式選取的方法。

結論

多維度存活資料的設限種類較一維問題更為多樣與複雜。如何針對不同的資料型態探討

變數間關聯性仍是熱門的課題。例如許多學者研究復發資料，此問題最大的特色是最後一階段受到依賴設限的問題，也有針對多階段病程模式的建構與其推論。此外如何將解釋變數的影響考慮在其中亦是實用的問題。有學者將此領域視為 longitudinal analysis 的一環。

參考文獻

- [1] W. Q. Meeker, *Statistical methods for reliability data*, Wiley, New York (1998).
- [2] E. L. Kaplan and P. Meier, *Journal of American Statistical Association*, **53**, 457; 562 (1958).
- [3] D. R. Cox, *Journal of Royal Statistical Society Series B*, **34**, 187 (1972).
- [4] O. O. Aalen, *Annals of Statistics*, **6**, 701 (1978).
- [5] T. R. Fleming and D. P. Harrington, *Counting Processes and Survival Analysis*, Wiley, New York (1991).
- [6] J. E. Anderson, T. A. Louis, N. V. Holm and B. Harvald, *Journal of American Statistical Association*, **87**, 641 (1992).
- [7] D. Oakes, *Journal of American Statistical Association*, **84**, 487 (1989).
- [8] C. Genest and L. P. Revist, *Journal of American Statistical Association*, **88**, 1034 (1993).
- [9] D. Clayton, *Biometrika*, **65**, 141 (1978).
- [10] R. D. Gill, *They of Probability and It's Applications*, **37**, 18 (1992).
- [11] D. M. Dabrowska, *Annals of Statistics*, **16**, 1475 (1988).
- [12] R. L. Prentice and J. Cai, *Biometrika*, **79**, 495 (1992).
- [13] R. C. Pruitt, *Annals of Statistics*, **19**, 443 (1991).
- [14] L. Hsu and R. L. Prentice, *Biometrika*, **83**, 905 (1996).
- [15] J. H. Shih and T. A. Louis, *Biometrics*, **51**, 1384 (1995).
- [16] W. Wang and M. T. Wells, *Journal of American Statistical Association*, **95**, 62; 75 (2000).