第一章 簡介

在臨床實驗上，我們把‘長期存活率’(long-term survival rate)視為癌症治癒的指標。例如：追蹤 5 年後，腫瘤是否有復發，若沒有復發，表示以後再發生的可能性很低，因此我們把‘長期存活率’當做癌症治好的指標。由此看來，估計‘長期存活率’顯得相當重要。

為什麼會有‘設限’資料產生呢？設限的種類很多，隨著領域的不同，所著重討論的設限情形也會不同。一般可分為型一設限(Type I censoring)、型二設限(Type II censoring)、隨機設限(Random censoring)以及其他型式的設限。型一設限和型二設限主要出現在工程應用上。隨機設限則較常發生在醫學的臨床實驗或動物研究上，本文所討論的設限即屬於隨機設限(Random censoring)。臨床實驗中，設限發生的原因大致有以下兩點：(1)假設我們訂某一個時間點為計畫結束的時間，但在計劃中途持續有病人進來，其就診時間會不滿此時間，因此會成為設限資料。(2)在計劃中途有病人轉診或是死於其他原因，無法繼續觀察結果，也會成為設限資料。

存活分析中，經常探討‘長期存活率 ’和一些因子之間的相關性。例如臨床實驗中研究τ天之存活率。若無病人的存活時間在τ天之內設限，我們可以利用典型的邏輯斯迴歸分析該資料。然而，大多

數的臨床實驗由於病人持續的就診，因此某些病人之存活時間在 τ 天之內設限。若將設限資料去除，則分析出現偏差。Jung(1996)提出一估計值 $\hat{b}$ 用以估計隨機設限資料下長期存活率邏輯斯迴歸之參數。第二章中，我們從分群的觀點探討 Jung 的方法。第三章中，利用 Rao and Scott (1992) 的方法，我們提出另一估計值 $\tilde{b}$ ，並證明 $\tilde{b}$ 比 $\hat{b}$ 漸近有效。此外，依據 $\tilde{b}$ 之卡方檢定可用於模式適合度檢定。我們以一臨床資料闡釋此方法。在第四章中，我們以模擬比較 $\hat{b}$ 和 $\tilde{b}$ 在有限樣本下之性質。最後在第五章做個討論。

第二章　Jung 的方法

令T_j（ j=1,2,….,n）表示個體 j 發生某一事件前所經歷的時間，例如：死亡、腫瘤復發等等。令Z_j^*為相對於個體 j 之已知共變數向量。假設族群依據Z^*區分為 k 個子族群(表示為 D_1，…..D_k)。 對於邏輯斯迴歸模型而言，‘長期存活率’ $p_i = P(T_i \geq t) = f_i(b)$ 滿足

$$\log[\ f_i(b)/\{1-\ f_i(b)\}] = z_i^T b \qquad (i = 1,....,\ k)$$

其中z_i為源於Z^*之維度 s ×1 已知共變數向量, β為維度 s ×1 未知迴歸參數向量。

若無病人的存活時間在 τ 天之內設限,我們可以利用典型的邏輯斯迴歸分析該資料。而 b 之最大概似解可經由下列遞迴解求得

$$Z^T D_n \hat{f} = Z^T D_n q \qquad (2.1)$$

其中 $Z^T = (z_{1,.....,} z_k)$ 為維度 s×k 已知共變數矩陣,其秩(rank)為 s， $\hat{f} = (f_1(\hat{b}),....., f_k(\hat{b}))^T$ $D_n = diag\ \{n_{1,.....,} n_k\}$

$q = (q_{1,.....} {}_{.,} q_k)^T$

其中 $q_i = n_{i1}/n_i$ ， n_i 為第 i^{th} 子族群之樣本數，

而 n_{i1} 為該子族群中存活時間超過 τ 天之個數。

對於設限資料,我們僅觀察到$(X_{j,} d_j)$ (j=1,2,..,n)

其中

$$X_j = \min(T_j, C_j) \quad , \quad d_j = I_{[T_j \le C_j]}$$

$C_1, \ldots, C_n$ 為設限時間。假設 $C_j's$ 為 iid 且和 $T_j's$ 獨立。

在每一子族群 $D_i (i = 1, \ldots, k)$ 中，假設設限時間之分配同為 $G_i(c) = P(C_j \le c \mid j \in D_i)$ 因無法觀察到所有 $T_j's$ ， 故 $n_{i1}'s$ 未知。Jung (1996)以 $\hat{n}_{i1} = \sum_{j \in D_i} I_{[X_j \ge t]} / [1 - \hat{G}_i(t)]$ 取代(2.1)中的 n_{i1}，其中 $\hat{G}_i(t)$ 為 $G_i(t)$ 之 Kaplan-Meier(1958)估計值。以下說明 $\hat{n}_{i1}$ 的推導：

(推導)：

$$E\{\sum_{j \in D_i} I(X_j \ge t \mid Z_j)\}$$

$$= \sum_{j \in D_i} P(X_j \ge t \mid Z_j) = \sum_{j \in D_i} P(T_j \ge t \mid Z_j) P(C_j \ge c \mid Z_j)$$

$$= \sum_{j \in D_i} S_i(t)[1 - G_i(t)] = n_i S_i(t)[1 - G_i(t)]$$

$$E\{\sum_{j \in D_i} I(X_j \ge t \mid Z_j) / 1 - G_i(t)\} = n_i S_i(t)$$

$$\therefore \hat{n}_{i1} = \sum_{j \in D_i} I_{[X_j \ge t]} / [1 - \hat{G}_i(t)]$$

其中 $\hat{G}_i(t) = 1 - \prod_{C_j \le t} (\frac{n_i - j}{n_i - j + 1})^{1 - d_j}$ (Kaplan-Meier 估計值)，

因 $\hat{n}_{i1} = n_i \hat{p}_i$ 其中 $\hat{p}_i$ 為 p_i 之 Kaplan-Meier 估計值，

Jung (1996)所提的估計值 $\hat{b}$ 相當於下列`擬'概度等式之遞迴參數

解:

$$Z^T D_n \hat{f} = Z^T D_n \hat{p}$$

其中 $\hat{p} = (\hat{p}_{1,.....,}\hat{p}_k)^T$ 和 $\hat{f} = (f_1(\hat{b}),....., f_k(\hat{b}))^T$ 。

令 w_i 為第 i^{th} 子族群個數佔總數的比例，

假設 $(\hat{n}_i / n) - w_i = o_p(1) \quad (i = 1,.....,k)$ ，則當

$n \to \infty, \sqrt{n}(\hat{p} - f(\boldsymbol{b}))$ 分配收斂致 $N(0, V_w)$ 。

其中 $V_w = diag\{V_1 / w_{1,.....} V_k / w_k\}$ ， V_i 為當 $n \to \infty$ 時，

$\sqrt{n_i}(\hat{p}_i - f_i(b))$ 的漸近變方。 $V_i = (\overline{G}_i(\boldsymbol{t}))^2 \int_0^t \frac{dG_i^*(t)}{[1 - H_i(t)]^2}$ ，

$\overline{G}_i(\boldsymbol{t}) = 1 - \overline{G}_i(\boldsymbol{t})$ ， $G_i^*(t) = P(C_j \le t, \boldsymbol{d}_j = 0 | j \in D_i)$ ， $H_i(t) = P(X_j \le t | j \in D_i)$ 。

依此， $\sqrt{n}(\hat{b} - b)$ 分配收斂致 $N(0, V_{\hat{b}})$ 。

其中 $V_{\hat{b}} = (Z^T \Delta Z)^{-1}(Z^T D_w V_w D_w Z)(Z^T \Delta Z)^{-1}$ ，

$D_w = diag\{w_1,.... w_k\}$ ， $\Delta = diag\{w_1 p_1(1 - p_1),.....,w_k p_k(1 - p_k)\}$ 。

下一章中，我們提出另一估計值 $\tilde{b}$ ，並證明 $\tilde{b}$ 比 $\hat{b}$ 漸近有效。

第三章 取代估計值

定義$c_i = V_i / p_i(1-p_i)]$為第i^{th} 子族群之`設限效應'。此$c_i$表示由於設限所導致的變方膨脹，其類似於抽樣的`設計效應'(參閱 Robert， Rao and Kumar 1987))。現在，以 $\hat{c}_i = \hat{V}_i / \hat{p}_i(1-\hat{p}_i)]$ 估計C_i。其中$\hat{V}_i = (1-\hat{G}_i(t))^2 \sum_{C_j \le t} \frac{1-d_j}{(n_i - j)(n_i - j+1)}$ (Greenwood (1926)之V_i估計值)。

定義 $\tilde{n}_i = n_i / \hat{c}_i = n_i[\hat{p}_i(1-\hat{p}_i)] / \hat{V}_i$為`有效樣本數'，

利用 Rao and Scott (1992)方法，以$\tilde{n}_i$ 和$\tilde{n}_{i1} = \tilde{n}_i \hat{p}_i$

取代(2.1)中的n_i 和 n_{i1}。

另一估計值為 $\tilde{b}$ `擬'概度等式之遞迴參數解:

$$Z^T D_{\tilde{n}} \tilde{f} = Z^T D_{\tilde{n}} \hat{p}$$

其中 $D_{\tilde{n}} = diag\ \{\tilde{n}_{1,.....,}\ \tilde{n}_k\}$ ， $\tilde{f} = (f_1(\tilde{b}),....., \ f_k(\tilde{b}))^T$ 。

因$\hat{V}_i = V_i + o_p(1), (i = 1,...., \ k)$ ，

當$n \to \infty$時， $\sqrt{n}(\tilde{b} - b)$ 分配收斂致 $N(0, V_{\tilde{b}})$。

其中$V_{\tilde{b}} = (Z^T \Delta_c Z)^{-1}$ ， $\Delta_c = diag\ \{\frac{w_1 p_1^2 (1-p_1)^2}{V_1},....., \ \frac{w_k p_k^2 (1-p_k)^2}{V_k}\}$

接下來，我們比較$\hat{b}$ 和 $\tilde{b}$ 。以下定理證明$V_{\hat{b}} - V_{\tilde{b}}$ 為半正定矩陣 (positive semidefinite)。

定理 3.1

$V_{\hat{b}} - V_{\tilde{b}}$ 為半正定矩陣 (positive semidefinite)。

證明:

令 $C = diag\ \{c_1,.., c_k\}, (i = 1,...., k)$ ， $c_i = V_i /[\ p_i(1-P_i)]$ ，則

$$V_{\hat{b}} = (Z^T \Delta Z)^{-1}(Z^T D_w V_w D_w Z)(Z^T \Delta Z)^{-1}$$

$$= (Z^T \Delta Z)^{-1}(Z^T C \Delta Z)(Z^T \Delta Z)^{-1} ,$$

$$V_{\tilde{b}} = (Z^T \Delta_c Z)^{-1} = (Z^T C^{-1} \Delta Z)^{-1}$$

我們僅需證明對於任一 $k \times 1$ 向量 y，

$$y^T[(Z^T\Delta Z)^{-1}(Z^T C\Delta Z)(Z^T\Delta Z)^{-1} - (Z^T C^{-1}\Delta Z)^{-1}]y \geq 0 \qquad (3.1)$$

令 $u = \Delta Z (Z^T \Delta Z)^{-1} y$ ，則 $u^T = y^T (Z^T \Delta Z)^{-1} Z^T \Delta$ 。

(3.1)式的第一項為

$$y^T[(Z^T\Delta Z)^{-1}(Z^T C\Delta Z)(Z^T\Delta Z)^{-1}]y$$

$$= y^T (Z^T\Delta Z)^{-1} Z^T \Delta\Delta^{-1} Cu = u^T \Delta^{-1} Cu = u^T C \Delta^{-1} u 。$$

(3.1)式的第二項為

$$y^T[(Z^T C^{-1}\Delta Z)^{-1}]y$$

$$= y^T (Z^T\Delta Z)^{-1}(Z^T\Delta Z)(Z^T(\Delta^{-1}C)^{-1}Z)^{-1}(Z^T\Delta Z)(Z^T\Delta Z)^{-1} y$$

$$= u^T Z (Z^T (C\Delta^{-1})^{-1} Z)^{-1} Z^T u 。$$

則(3.1) 相當於

$$u^T [C\Delta^{-1} - Z(Z^T(C\Delta^{-1})^{-1}Z)^{-1}Z^T]u \geq 0 \qquad (3.2)$$

令 $C\Delta^{-1} = A^2, v = Au, H = A^{-1}Z$ ，則(3.2)相當於

$$v^T[I_{k\times k} - H(H^T H)^{-1}H^T]v \geq 0 \qquad (3.3)$$

$\because v^T v = u^T A^T Au = u^T A^2 u = u^T C\Delta^{-1}u$,

$$(H^T H)^{-1} = (Z^T A^{-1}A^{-1}Z)^{-1} = (Z^T (C\Delta^{-1})^{-1}Z)^{-1}$$

$$v^T[H(H^T H)^{-1}H^T]v = u^T AA^{-1}Z(Z^T (C\Delta^{-1})^{-1}Z)^{-1}Z^T A^{-1}Au$$

$$= u^T Z(Z^T (C\Delta^{-1})^{-1}Z)^{-1}Z^T u$$

∴(3.2)式 = (3.3)式。

其中$v^T[I_{k\times k} - H(H^T H)^{-1}H^T]v$ 為 v 對 H 迴歸之殘差平方和，其為半正定矩陣。故得證。

由定理 3.1，當$c_1 = ... = c_k = c$，我們得到 C =$cI_{k\times k}$且$V_{\tilde{b}} = V_{\hat{b}}$。

模式(2.1)之適合度檢定可經由以下卡方統計量(chi-squared statistic)

$$\tilde{X}^2 = \sum_{i=1}^{k} \tilde{n}_i(\hat{p}_i - f_i(\tilde{b}))^2 / f_i(\tilde{b})[1 - f_i(\tilde{b})]。$$

我們可證明在(2.1)模式下，$\tilde{X}^2$會漸近收斂到 X^2_{k-s} ，

其中 X^2_{k-s} 為自由度$k-s$之卡方變數。此證明類似沈葆聖(1998)定理 2.1，故省略。

接下來，我們以一退伍軍人肺癌臨床資料闡釋此方法(參閱

Prentice (1973))。一百三十七個病人隨機給予試驗性的化療，我們

		$\hat{b}$			$\tilde{b}$		
效應	df	估計值	std	p-value	估計值	std	p-value
intercept	1	-4.179	0.934	0	-2.683	0.886	0.003
ADN	1	-2.483	1.001	0.013	-2.199	0.709	0.002
SM	1	-1.442	0.594	0.015	-1.328	0.575	0.021
PS	1	0.038	0.012	0.002	0.033	0.012	0.006
SQM	1	0.509	0.472	0.281	0.35	0.541	0.517
Goodness-of-fit	11			0.344			0.322

考慮兩個共變數，外科手術狀況(PS)和腫瘤型態(鱗狀型(squamous)，小型(small)或腺型(adeno))。t 值設定為 180 天。9 筆資料為設限資料，其中 8 筆在 180 天前設限。 我們考慮下列邏輯斯迴歸模式:

$$\log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 PS_i + \beta_2 SQM_i + \beta_3 SML_i + \beta_4 ADN_i \text{，} \quad (i = 1,.....,12)$$

如為鱗狀型則 $SQM_i = 1$，否則 $SQM_i = 0$；如為小型則 $SML_i = 1$，否則 $SML_i = 0$；如為腺型則 $ADN_i = 1$，否則 $ADN_i = 0$；$PS_i = 25$ 對於 $PS \le 30$，$PS_i = 45$ 對於 $40 \le PS \le 50$，$PS_i = 65$ 對於 $60 \le PS \le 70$，$PS_i = 85$ 對於 $PS \ge 80$。族群依據外科手術狀況和腫瘤型態區分為 12 個子族群。我們計算 $\hat{b}$ 和 $\tilde{b}$ 估計值。表 1 列出迴歸參數估計，自由度(df)，標準差(std)，以及適合度檢定之 p-values。

表 1． $\hat{b}$ 和 $\tilde{b}$ 之邏輯斯迴歸分析； t =180 天

表 1 顯示 ADN，SM 和 PS 皆顯著(注意到依據 $\hat{b}$ 所計算之 p-value 是不正確的，而依據 $\tilde{b}$ 所計算之 p-value 是正確的，為 0.322)。由於依據 $\tilde{b}$ 所計算的 p-value=0.322，表示此邏輯斯迴歸模式是合適的。

第四章 模擬結果

本章中，我們以模擬比較 $\hat{b}$ 和在 $\tilde{b}$ 有限樣本下的表現。樣本數 n 設為 100, 當 $j \le 25$，共變 Z_j 設為 0, 當 25< $j \le 50$ 設為 1, 當 50< $j \le 75$，設為 2，當 j > 75 設為 3。$T_j (j = 1,....,n)$ 為指數分配 (exponential distribution) 具期望值 1。在此設定下，我們考慮 $p_i = P(T_j \ge 1 | Z_j = i)$ $(i = 0,1,2,3)$ 之邏輯斯迴歸模式：

$$\log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 Z_i$$

β_1真正值設為 0，而β_0真正值設為$\log(e^{-1}/[1-e^{-1}]) = -0.541$。設限變數為均勻分配(uniform distributions) $U(dZ_{i,}1+dZ_i)$。選擇d =0.2，0.4，0.6 和 0.8 依此函蓋嚴重到輕微設限。模擬重覆 10000 次。表 2 顯示兩估計值之設限比例(表為p^*)，偏差(biases)，標準差(std)和均方差(mean squared errors (mse))。

表 2．$\hat{\beta}_0$和$\tilde{\beta}_0$之模擬結果

		Bias		Std		Mse	
d	p^*	$\hat{\beta}_0$	$\tilde{\beta}_0$	$\hat{\beta}_0$	$\hat{\beta}_0$	$\hat{\beta}_0$	$\tilde{\beta}_0$
0.2	0.395	-0.018	0.102	0.766	0.678	0.587	0.471
0.4	0.258	-0.014	0.018	0.61	0.586	0.372	0.343
0.6	0.175	-0.015	-0.006	0.557	0.548	0.31	0.301
0.8	0.123	-0.017	-0.015	0.527	0.525	0.278	0.276

p^*：設限比例(根據模擬計算而得到的值)

表 3． $\hat{b}_1$和 $\tilde{b}_1$之模擬結果

		Bias		Std		Mse	
d	p^*	$\hat{b}_1$	$\tilde{b}_1$	$\hat{b}_1$	$\hat{b}_1$	$\hat{b}_1$	$\tilde{b}_1$
0.2	0.395	0.003	-0.029	0.251	0.228	0.064	0.053
0.4	0.258	0.002	-0.007	0.212	0.205	0.045	0.042
0.6	0.175	0.002	-0.001	0.197	0.195	0.039	0.038
0.8	0.123	0.003	0.002	0.19	0.19	0.036	0.036

從表 2 和表 3，我們觀察到當嚴重設限時(d=0.2)，$\tilde{b}$ 的偏差大於 $\hat{b}$。然而，$\tilde{b}$ 的標準差小於 $\hat{b}$ 且隨著設限比例增加，$\tilde{b}$ 的效率亦增加。 所有模擬考慮的狀況，$\tilde{b}$ 的均方差皆小於 $\hat{b}$。

第五章 討論

本文所提出的估計值 $\tilde{b}$，雖然可做為模式適合度檢定之依據，但僅適用於共變數為分類資料。此外，由於有效樣本涉及 Kaplan-Meier 變方估計，因此，在樣本數不夠大的情況下，$\tilde{b}$ 未必優於 $\hat{b}$ 。後續的研究應著重於：如何在共變數為連續資料下，診斷模式的合適與否。