

GEE 長期追蹤資料分析在資料缺失下的模型選取問題

程毅豪

中央研究院統計科學研究所

一、簡介

長期追蹤研究 (longitudinal study) 廣泛出現於社會學、生物學及醫學等學科的研究領域。在長期追蹤研究中，每個研究對象在多個時間點被觀察，由此長期追蹤所得到數據中，研究者可了解所感興趣的應變量 (如體重、疾病狀態、事件發生次數等) 如何隨時間及其他因變量 (如運動頻率、用藥類別、初始事件次數) 而變化，亦可了解不同時間點應變量間的相關情況。在長期追蹤資料分析的統計方法中，由 Kung-Yee Laing (梁賡義) 與 Scott Zeger 於 1986 年所提出的 generalized estimating equations (GEE) 方法是一廣被使用的方法，它可用來分析各時間點之應變量與因變量間的關係，即長期追蹤資料的邊際迴歸分析 (marginal regression analysis)。GEE 分析的特點在於，其考慮了各時間點應變量間的相關結構，且其迴歸參數的正確估計 (consistent estimation) 在不正確的應變量間相關結構假設下仍然成立，因此具有統計推論上的穩健性 (robustness)。

在迴歸分析中，迴歸模式的選取 (model selection) 一直是重要的研究與實務問題。有關 GEE 的模型選取問題，亦有若干文獻加以討論，包括 Pan (2001) 及 Cantoni, Mills Flemming, and Ronchetti (2005)。其中 Pan (2001) 所提出的 Quasi-likelihood under the Independence model Criterion (QIC) 方法已被納入常用的統計軟體 (如 SAS, R 等) 而廣被使用，其可視為 Akaike Information Criterion (AIC) 方法在長期追蹤資料的延伸。但在長期追蹤資料中，有一重要因素必須考慮，即長期追蹤資料常見缺失資料 (missing data)，而現存有關於 GEE 分析的模型選取方法並未考慮此因素。事實上，缺失資料對統計分析的影響，文獻中的討論大都集中於參數估計問題，即缺失資料可造成參數估計的偏誤以及參數估計變異或標準誤的增加。但關於缺失資料對模型選取的影響，文獻中則尚未有較系統化的討論與研究。因此，如何在有缺失資料的情形下對 GEE 模型進行有效的模型選取，變成為實務上相當重要，但方法學上尚未有成熟解答的問題。

二、Missing Longitudinal Information Criterion

有鑑於此，本人與中正大學統計科學研究所的沈仲維副教授於 2012 年發表了針對在應變量有缺失資料下 GEE 分析的模型選取方法 (Shen and Chen, 2012)。在此項研究中發現，在應變量資料有缺失的情形中，忽略缺失資料而直接將觀察到的資料視為完整長期追蹤資料，並套用 QIC 方法進行模型選取，所選到的模型往往偏離正確的模型。例如，考慮應變量為二元資料的長期追蹤資料，每個體長期觀察次數為 3，應變量在不同時間點的相關係數為一常數 0.25，即相關結構屬“compound symmetry” (CS)。真實的邊際迴歸模型為包含兩個自變量 X_1 、 X_2 的 logistic model，其中 X_1 為與不隨時間變化的二元變數， X_2 為觀測時間。另有兩個自變量 X_3 、 X_4 為互相獨立且不隨時間變化的均勻分布隨機變量，其為與應變量無關的多餘自變量。部分的應變量資料有缺失，缺失比例 (missing proportion) 為 0%、20% 或 35%。在 1000 次模擬中，我們計算 5 個候選模型 (如表一所示) 被 QIC 模型選取方法選中的百分比。由表一中可看出，當自變量有缺失時 (missing proportion = 20, 35%)，QIC 方法選中正確模型 $\{X_1, X_2\}$ 的比例偏低，且有超過 50% 的頻率會選中錯誤的模型 $\{X_1\}$ 而忽略掉另一個與應變量有關的自變量 X_2 。

表一. MLIC(CS), MLIC(Ind),及 QIC 方法中,各候選模型被選取的百分比(1000 次模擬中)。正確模型為 $\{1,2\}=\{X1,X2\}$ 。MLIC(CS) 與 MLIC(Ind)分別為在 compound symmetry 與 independence 相關結構假設下計算之 MLIC 統計量。

N	Missing Proportion	Method	Model				
			{1}	{1,2}*	{1,3}	{1,2,3}	{1,2,3,4}
50	0%	MLIC(CS)	35.4	41.3	6.6	7.2	9.5
		MLIC(Ind)	35.4	41.3	6.7	7.1	9.5
		QIC	37.5	38.7	6.9	7.4	9.5
	20%	MLIC(CS)	25.9	33.9	12.3	11.0	16.9
		MLIC(Ind)	25.9	34.2	12.3	11.0	16.6
		QIC	59.5	15.4	13.9	3.2	8.0
	35%	MLIC(CS)	27.0	28.4	12.6	11.7	20.3
		MLIC(Ind)	27.0	28.7	12.7	11.5	20.1
		QIC	65.5	11.7	13.6	2.7	6.5
100	0%	MLIC(CS)	22.5	56.2	2.7	9.9	8.7
		MLIC(Ind)	22.5	56.3	2.7	9.8	8.7
		QIC	22.9	55.5	2.8	9.7	9.1
	20%	MLIC(CS)	19.6	43.3	6.4	13.4	17.3
		MLIC(Ind)	19.6	43.4	6.5	13.2	17.3
		QIC	56.7	19.4	12.2	3.2	8.5
	35%	MLIC(CS)	17.7	37.5	9.8	12.3	22.7
		MLIC(Ind)	17.7	37.7	9.8	12.2	22.6
		QIC	64.3	15.1	11.5	2.6	6.5

*: 正確模型

上述的例子說明了缺失資料對 GEE 分析的模型選取有不可忽略的影響：忽略此影響而逕自套用現有未考慮缺失資料的模型選取方法將造成模型選取的偏誤。我們因此重新推導了在有缺失資料下，正確的模型選取準則統計量。考慮某候選模型的期望平方誤差，亦即該候選模型與真實模型的差距。此誤差數量越小代表該模型是愈“好”的模型，反之則代表愈“差”的模型。經過計算，該誤差為兩個數量之總和，其一為應變量與候選模型的差距，即該候選模型的適合度 (goodness of fit)；另一為應變量與候選模型間的共變異數 (covariance)，其反映了該候選模型與應變數資料的相依程度，即模型的複雜度 (complexity) 或過度配適 (overfitting) 的程度，又稱為“covariance penalty” (Efron, 2004)。此兩數量反映的

是一個模型在配適資料時所面臨的兩難抉擇 (dilemma): 較“大”(自變量個數較多)的模型通常有較好的適合度但有較大的複雜度,而較“小”(自變量個數較少)的模型通常有較差的適合度但有較小的複雜度。因此若以期望平方誤差為模型選擇的標準,則一個最“好”(期望平方誤差最小)的模型需在適合度與複雜度間取一個最適當的平衡。

Shen and Chen (2012) 推導了在有缺失應變量資料的 GEE 分析中,適合度數量 (goodness of fit) 與複雜度數量 (covariance penalty) 的近似不偏估計量,從而得到期望平方誤差的近似不偏估計量,並建議以此統計量作為 GEE 模型選取的準則,我們稱之為 Missing Longitudinal Information Criterion (MLIC)。由其定義可知,一個愈“好”的模型有愈小的 MLIC 值,即在所有候選模型中,具有最小 MLIC 數值的模型將是被選中的模型。關於 MLIC 統計量的詳細計算公式與推導,請參見 Shen and Chen (2012, equation 6)。

表一呈現了在上述模擬研究中,以 MLIC 為準則的模型選取結果,其中 MLIC(CS) 與 MLIC(Ind) 分別為在 compound symmetry 與 independence 相關結構假設下計算之 MLIC 統計量。我們可以看到,MLIC 方法在有缺失資料的情形中 (missing proportion=20, 35%), 仍然有相當高的頻率 (>33%) 選中正確的模型 {X1,X2}。同時,不同的相關結構假設 (compound symmetry, independence) 對 MLIC 模型選取結果的影響甚小,意味著 MLIC 保留了 GEE 方法對相關結構假設的穩健性。

三、實例分析

表二呈現了我們分析一組來自 National Institute of the Mental Health Schizophrenia Collaborative Study 的結果 (Shen and Chen, 2012)。在此資料中,應變量資料為386位精神分裂症病人在4個時間點的病情狀態(以二元資料呈現病情嚴重或不嚴重),其中有77位病人因中途離開研究而有缺失資料。自變量資料則包括這些病人的用藥 (Drug; 1: 處方藥; 0: 安慰藥)、性別 (Sex; 1: 男性; 0: 女性)、及時間 (Time)。在應變數追蹤資料具有自我相關結構 (AR-1) 的假設下,我們考慮如表二所列的5個候選GEE迴歸模型,並計算其對應之 MLIC 及 QIC 模型選取準則統計量。由表二,MLIC方法選擇了 model 2 (包含時間與用藥為自

變數的 logistic model)，而QIC 方法則選擇了 model 4 (包含時間、用藥及其交互作用為自變量的 logistic model)。由於時間與用藥在 model 2 均達到統計上的顯著性 ($p \text{ value} < 0.05$)，而其交互作用項在 model 4 並未達到統計上的顯著性 ($p \text{ value} > 0.05$)，因此MLIC所選出的模型 (model 2) 較 QIC 所選出者 (model 4) 似乎更為合理 (模式中所有自變數均具備統計顯著性)。

表二. National Institute of the Mental Health Schizophrenia Collaborative Study 資料分析結果: 參數估計 $\pm$ 標準誤，MLIC 與 QIC 統計量數值。

covariate	Model				
	1	2	3	4	5
Time(T)	-1.34 $\pm$ 0.08*	-1.37 $\pm$ 0.08*	-1.37 $\pm$ 0.08*	-1.16 $\pm$ 0.21*	-1.18 $\pm$ 0.24*
Drug(D)		-0.85 $\pm$ 0.24*	-0.86 $\pm$ 0.24*	-0.35 $\pm$ 0.44	-0.52 $\pm$ 0.49
Sex(S)			0.12 $\pm$ 0.18		-0.18 $\pm$ 0.49
T $\times$ D				-0.26 $\pm$ 0.23	-0.25 $\pm$ 0.23
T $\times$ S					0.02 $\pm$ 0.17
S $\times$ D					0.34 $\pm$ 0.46
MLIC(AR-1)	261.9	255.7	256.5	255.9	257.5
QIC(AR-1)	1406.3	1372.7	1374.9	1372.4	1378.8

*: $p \text{ value} < 0.05$

四、結語

我們所提出的長期追蹤資料分析模型選取方法，是針對資料被不完全觀測情形下，迴歸模型選取問題的系統性研究。此方法學可進一步推廣至更一般的資料缺失型態 (Shen and Chen, 2013)，當自變量資料同時有資料缺失及測量誤差的情形(Shen and Chen, 2015)，以及同時對迴歸模型與相關結構模型進行模型選取的問題(Shen and Chen, 2018)。對於長期追蹤研究中普遍存在的缺失或測量誤差資料，這些統計方法學提供了有效的模型選取工具。

參考文獻

- [1] C.W. Shen, Y.H. Chen. Biometrics, 68, 1046 (2012).
- [2] C.W. Shen, Y.H. Chen. Biometrical Journal, 55, 899 (2013).
- [3] C.W. Shen, Y.H. Chen. Biostatistics, 16, 740 (2015).
- [4] C.W. Shen, Y.H. Chen. Biometrical Journal, 60, 20 (2018).
- [5] B. Efron. Journal of the American Statistical Association, 99, 619 (2004).