一葉報秋: 互訊息在類別資料分析的應用

劉長萱研究員（統計科學研究所）

＊特別感謝程爾觀、呂玉瑕及高鴻文教授的指正及建議。

一、簡介

多變量類別(categorical)資料分析在社會科學及醫療保健研究領域，佔了相當大的比例；例如採用類別資料探討親子互動與學齡兒童情緒問題的關聯。類別資料可採多向度列聯表(contingency table)方式表示，表中呈現不同屬性同時出現的頻數或比例；例如經常使用手機的男性並罹患腦瘤在總樣本中佔的比例。「對數線性」模式(log-linear model)，及與該模式密切相關的「勝算對數」模式(logit model) 為一般分析列聯表的統計方法，兩種方法都可用概似比(likelihood ratio)檢定模式估計值與實際觀察值是否接近。概似比值在統計分析軟體 (例如SAS, SPSS)中又稱為離差值(deviance)，該值之統計顯著性越強代表「模式」對資料的解釋力越弱，應用時可採離差值考驗模式假設。傳統類別資料分析裡，較少參考訊息理論(information theory)中的基本定理或法則，兩者實際上有密不可分的關係；例如兩個變數的「相互概似比」在訊息理論中又稱為互訊息(mutual information；Kullback & Leibler, 1951)；此處簡稱MI。訊息理論的應用不僅可擴展對數線性模式及勝算對數模式的應用(例如考慮有用的高階交互作用)，也可協助重新檢視文獻中模式應用的方式是否正確。此處介紹一個訊息恆等式法則(Pythagorean Law of information identity; Cheng, Liou, Aston, Tsai, 2008)，並以實際數據介紹該法則的應用。

假設X, Y, 及Z為三個類別變數，I(X; Y; Z) 定義為變數之間的互訊息量；該訊息量的估計可在統計分析軟體中，選擇對數線性模式{X, Y, Z} (此處採Agresti，2013 使用的模式符號)；也就是只包含主效果的模式，軟體輸出該模式的離差或概似比值即為三個變數的互訊息量。該互訊息量可以拆解成下列三個正交的訊息成分 (Cover & Thomas, 1991):

$$I(X; Y; Z) = I(X; Z) + I(Y; Z) + I(X; Y|Z)。$$

恆等式右邊 I(X; Z)及I(Y; Z) 為二維的互信息，而I(X; Y|Z)為三維的條件互信息(CMI; conditional mutual information)；因此，三個變數的MI可以拆解為2個二維MI及1個三維CMI。恆等式中條件互信息等同於對數線性模式{XZ, YZ}的離差值；也就是模式中包含三個主效果及XZ、YZ的交互作用，該值可進一步拆解成:

$$I(X; Y|Z) = Int(X; Y; Z) + Par(X; Y|Z)。$$

式中Int(X; Y; Z)為三個變數的交互作用，可採模式{XZ, YZ, XY}的離差值估計。Par(X; Y|Z)為{X, Y}在控制分層變數Z後的淨關聯值(partial association)；由於正交性質，該值可用I(X; Y|Z) 減去 Int(X; Y; Z)的剩餘差值估計。由於三個變數只容許兩個二維的MI，所以若XZ, YZ已在模式中，XY對離差值的實際貢獻為三維的Par(X; Y|Z)，而非二維的交互作用；所以不參考訊息法則，很難解讀XY在模式中的實質作用。醫學界常使用的Cochran-Mantel-Haenszel (CMH)檢定，原理上等同檢定淨關聯Par(X; Y|Z)的樣本估計值；因此，若樣本估計的Int(X; Y; Z)值顯著時，則無需檢定「淨關聯」或CMH值。總的來說，「兩階段檢定法」先檢定估計的Int(X; Y; Z)值是否顯著，若不顯著再檢定兩個變數X及Y在控制Z後的「淨關聯」值是否顯著(Cheng, Liou & Aston, 2010)。四個變數的MI可以拆解為3個二維的MI、及2個三維和1個四維的CMI；以下採實例說明該訊息恆等式的應用。

二、實例

表一列聯表中病例組為國內某教學醫院缺血性中風(ischemic stroke)病患，對照組為無明顯腦傷病患(Kao et al., 2015)，兩組受試者皆簽署了參與研究同意書。例子中血壓、血糖、及年齡為解釋變數，缺血性中風為主要預測變數。社會科學或醫療保健研究中，經常使用僅含主效果的勝算對數模式:

$$\text{logit}[\, f(\text{中風}=1|\text{血壓}=j,\ \text{血糖}=k,\ \text{年齡}=l)\,] = \log\left(\frac{f_{1,jkl}}{f_{0,jkl}}\right) = \beta_0 + \beta_j^{\text{血壓}} + \beta_k^{\text{血糖}} + \beta_l^{\text{年齡}},$$

式中$f_{i,jkl}$代表四個變數在i、j、k、l屬性或層級的比例(密度函數)，β則為模式參數。勝算對數模式中已自動納入血壓、血糖、與年齡三個解釋變數的MI，該MI值已含了2個二維的MI及1個三維的CMI。根據訊息恆等式法則，四個變數的互訊息量扣除三個解釋變數的MI後，剩餘的訊息量為: (1) I(中風；血壓) ， (2) I(中風；血糖|血壓) ，及 (3) I(中風；年齡|{血糖，血壓})；應用時三個解釋變數的次序，可依需要調整。我們先在迴歸模式加入一個解釋變數「血壓」，該變數對降低概似比值的貢獻為二維的MI；也就是 (1)中的 $\hat{I}$ (中風; 血壓) = 105.425 (自由度為1, 顯著性p < .001)；此處 $\hat{I}$ (·)為 I(·)的樣本估計值。接著迴歸模式中再加入第二個解釋變數「血糖」，該變數對降低概似比值的貢獻為(2)中的$\widehat{\text{Par}}$(中風; 血糖｜血壓)=5.238 (自由度為1, 顯著性p = .022)。現在四個變數的總訊息量僅剩(2)中的Int(中風; 血糖|血壓)及(3)中的I(中風；年齡|{血糖，血壓})需估計。當此模式中再加入第三個解釋變數「年齡」，該變數對降低概似比值的貢獻量為39.609 (自由度為1, 顯著性p < .001)，但該值並沒有對應的MI或CMI可參照；也就是應用上無法正確解讀軟體提供的Type-I概似比值39.609的實質意義。在加入「年齡」之前，模式中應該先加入血壓及血糖的交互作用；該交互作用對概似比值的貢獻為(2)中的$\widehat{\text{Int}}$(中風; 血壓；血糖)=13.571 (自由度為1, 顯著性p < .001)。最後加入解釋變數「年齡」，此時該變數對概似比值的貢獻為(3)中的 $\widehat{\text{Par}}$(中風; 年齡|{血壓,血糖}) = 35.122 (自由度為1, 顯著性p < .001)，且該整體模式(3個主效果及1個交互作用)的離差值為(3)中的$\widehat{\text{Int}}$(中風; 年齡；{血壓,血糖}) = 23.484 (自由度為3, 顯著性p < .001)。此實例說明勝算對數模式僅考慮解釋變數的主效果，無法滿足一個有效的訊息恆等式，也無法正確評估解釋變數在該模式中的實質貢獻；模式加入一個三維的交互作用後，便可準確地估計「年齡」與「中風」的淨關聯值。預測缺血性中風的實用模式，仍需考慮原資料檔中其他解釋變數，表一中三個解釋變數僅用來說明互訊息的用途。上述模式的離差值23.484較大，代表模式配適的列聯表和觀察的列聯表(表一)顯著不同(p < .001)。

表一　缺血性中風的危險因子數據

			缺血性中風		
血壓	血糖	年齡	控制組	病患組	Total
正常	正常	≤ 60	482	16	498
		> 60	246	45	291
		Total	728	61	789
	FSG > 7.8mmol/L	≤ 60	38	6	44
		> 60	62	21	83
		Total	100	27	127
SBP >140mmHg或	正常	≤ 60	132	23	155
DBP >90mmHg		> 60	299	140	439
		Total	431	163	594
	FSG > 7.8mmol/L	≤ 60	58	27	85
		> 60	201	76	277
		Total	259	103	362

三、結論

恆等式法則係根據幾何的正交性質拆解變數間的互訊息量，此與代數估計主效果或交互作用參數的概似比值不盡相同；例如前述勝算對數模式的參數概似比值，與相對應的幾何MI值並不等價。在滿足訊息恆等式的情況下，幾何與代數途徑相似處為配適的列聯表(fitted table)相同，Type-I概似比與對照的MI或CMI相同，及最後進入模式的解釋變數$\widehat{\text{Par}}(\cdot)$值與對應的Type-III概似比值相同。相較於傳統的Akaike訊息準則，訊息理論對類別資料分析的貢獻在於選擇解釋變數之間的交互作用，也就是參考$\widehat{\text{Int}}(\cdot)$值。資料分析中，常遇到解釋變數的$\widehat{\text{Par}}(\cdot)$值不顯著，但$\widehat{\text{Int}}(\cdot)$值顯著；例如實例中$\widehat{\text{Par}}$(中風; 血糖|{血壓,年齡 }) = 2.678 (自由度為1, 顯著性p = .102)，但$\widehat{\text{Int}}$(中風; 血糖|{血壓,年齡 }) = 19.690 (自由度為3, 顯著性p < .001)。建立勝算對數模式時，若依據$\widehat{\text{Par}}(\cdot)$值篩選解釋變數，將忽略「血糖」對預測「缺血性中風」的潛在貢獻。並非所有考慮交互作用的勝算對數模式皆能滿足一個有效的訊息恆等式，不滿足恆等式的情況下，參數解讀將遇到類似前述主效果模式的困難；建立模式需參考並檢定$\widehat{\text{Int}}(\cdot)$值，並在加入主效果及交互作用同時，參考訊息恆等式法則。

四、參考資料

1. Agresti, A. (2013). Categorical Data Analysis (3nd Ed). New Jersey: Wiley.
2. Cheng, P. E.; Liou, M., Aston J. A. D., & Tsai, A.C. (2008). Information Identities and Testing Hypotheses: Power Analysis for Contingency Tables. Statistica Sinica, 18, 535-558.
3. Cheng, P. E.; Liou, M., & Aston J. A. D. (2010). Likelihood ratio tests with 3-way tables. Journal of the American Statistical Association, 105, 740-749.
4. Cover, T. M. & Thomas, J. A. (1991). Elements of Information Theory. New York: Wiley.
5. Kao, H. W.; Liou, M.; Chung, H. W.; Liu, H. S.; Tsai, P. H.; Chiang, S. W.; Chou, M. C; Peng, G. S.; Huang, G. S.; Hsu, H. S. & Chen, C. Y. (2015). Middle cerebral artery calcification association with ischemic stroke. Medicine 94 (50): e2311.