[研究成果報導]

基因與環境因子的交互作用之病例分析

中國醫藥大學公共衛生學系　鄭光甫　林維鈞

基因與環境因子的交互作用(gene-environment interactions)對疾病的影響一直是流行病學家(epidemiologist)有興趣的研究目標。由於對照組(controls)的基因資料較難收集，所以流行病學家需要使用最少的對照組資料即能對交互作用做精準的統計推論的方法。針對此目的，許多流行病學家建議不要使用對照組的資料，只使用病例組的資料對交互作用做統計推論，此種研究設計稱爲病例研究設計(case-only design)。雖然病例研究設計被廣泛的運用在交互作用的統計推論上，但是基因型判定錯誤(genotyping error)會使得此類方法所得到的結論產生偏差(bias)，而在現實資料中，基因型判定錯誤是難以避免的。在病例研究設計下，我們提出一種最大概似法(maximum likelihood method)對基因與環境因子的交互作用做統計推論，此方法的運用等於配適一多項式邏輯斯模型(multinomial logistic model)，使用上非常方便。此外，在病例研究設計下，我們利用基因型重複判定的觀念提出一種能有效修正基因型判定錯誤造成的偏差的方法，此方法也是最大概似法同時不需要事先知道基因型判定錯誤模型(genotyping error model)。

一、基因與環境因子的交互作用之統計推論

1. 病例研究之最大概似法

令 D 代表生病的狀況，1 表示生病（病例組），0 表示健康（對照組）。假設有 K 個基因因子(genetic factor)和 L 個環境因子(environmental factor)，且在對照組中，這些基因因子 $G = (G_1, \ldots, G_K)'$和環境因子 $E = (E_1, \ldots, E_L)'$在給定變數 W 時條件獨立。我們假設 $W = (W_0, \ldots, W_M)'$爲有$(M + 1)$個類別的類別變數(categorical variable)，譬如階層變數(stratification variable)，其中 W_m 等於 1 代表基因與環境因子屬於第 m 個類別，等於 0 爲其他情形，而 $W = 0$ 代表基因與環境因子互相獨立。此節的結果是推廣自 Piegorsch et al. [1]、Begg & Zhang [2]、Umbach & Weinberg [3]、Albert et al. [4]和 Armstrong [5]的結果。

假設我們有一組來自病例組且樣本數爲 n 的隨機樣本(random sample) $\{g_i, e_i, w_i\}$, $i = 1, 2, \ldots, n\}$，以下我們將利用這組樣本推導交互作用的最大概似法的統計推論。爲了使式子看起來較簡明，我們假設基因因子爲二項因子(binary factor)，等於 1 代表樣本帶有高風險的基因型，等於 0 代表樣本帶有低風險的基因型，然而基因因子可爲任意個類別，此時的做法是類似的。環境因子可以是類別的或是連續的。我們給定一個邏輯斯模型爲

$$\begin{aligned}&\operatorname{logit}\{P(D=1\mid G=g,E=e,W=w)\}\\&=\alpha^{*}+\sum_{k=1}^{K}\beta_k g_k+\sum_{l=1}^{L}\gamma_l e_l+\sum_{m=1}^{M}\psi_m w_m+\sum_{k=1}^{K}\sum_{r=k+1}^{K}\phi_{kr}g_k g_r\\&\quad+\sum_{l=1}^{L}\sum_{s=l+1}^{L}a_{ls}e_l e_s+\sum_{k=1}^{K}\sum_{l=1}^{L}\theta_{kl}g_k e_l\\&\quad+\sum_{k=1}^{K}\sum_{m=1}^{M}b_{km}g_k w_m+\sum_{l=1}^{L}\sum_{m=1}^{M}c_{lm}e_l w_m\\&\quad+\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}d_{klm}g_k e_l w_m\end{aligned}$$

其中 ϕ_{kr} , a_{ls} , θ_{kl} , b_{km} 和 c_{lm} 爲兩因子的交互作用，d_{klm} 爲三因子的交互作用。在此只有交互作用θ_{kl}和d_{klm}爲可估計的。

接下來，爲了展現輪廓概似函數(profile likelihood function)，我們定義 S 爲基因因子所有可能的組合的集合(set)且令 $t = (t_1,\ldots,t_K)'$ 爲集合 S 的任意元素(element)。我們進一步定義

$$\begin{aligned}\beta^{*}(t)&=\sum_{k=1}^{K}\beta_k t_k+\sum_{k=1}^{K}\sum_{r=k+1}^{K}\phi_{kr}t_k t_r\\&\quad+\ln\{f_0(G=t\mid W_0=1)\}\end{aligned}$$

$$\psi_m^{*}=\psi_m+\ln\left\{\frac{f_0(G=0\mid W_m=1)}{f_0(G=0\mid W_0=1)}\right\}$$

$$\theta_l(t)=\sum_{k=1}^{K}\theta_{kl}t_k$$

$$b_m^*(t)=\sum_{k=1}^{K}b_{km}t_k+\ln\left\{\frac{f_0(G=t\,|\,W_m=1)f_0(G=0\,|\,W_0=1)}{f_0(G=0\,|\,W_m=1)f_0(G=t\,|\,W_0=1)}\right\}$$

$$d_{lm}(t)=\sum_{k=1}^{K}d_{klm}t_k$$

其中 $f_0(G=t\,|\,W_m=1)$ 爲在對照組中給定 W_m = 1 時，G = t 的條件機率。則我們可以證明去除常數項的輪廓概似函數爲

$$L_p^*=\prod_{i=1}^{n}\frac{\sum_{t\in S}I(g_i=t)\exp\left\{\beta^*(t)+\sum_{l=1}^{L}\theta_l(t)e_{il}+\sum_{m=1}^{M}b_m^*(t)w_{im}+\sum_{l=1}^{L}\sum_{m=1}^{M}d_{lm}(t)e_{il}w_{im}\right\}}{\sum_{t\in S}\exp\left\{\beta^*(t)+\sum_{l=1}^{L}\theta_l(t)e_{il}+\sum_{m=1}^{M}b_m^*(t)w_{im}+\sum_{l=1}^{L}\sum_{m=1}^{M}d_{lm}(t)e_{il}w_{im}\right\}}$$

其中 $I(g_i=t)$ 爲指標函數(indicator function)。由輪廓概似函數可得知 θ_{kl} 和 d_{klm} 的最大概似估計值可利用病例組的資料 $\{(g_i,e_i,w_i),i=1,2,...,n\}$ 並配適一個廣義的多項式邏輯斯迴歸模型而得到，而變異數的估計值爲觀察的訊息矩陣(observed information matrix)的反矩陣。而諸如概似比檢定(likelihood ratio test)等建立在概似函數下的檢定統計量也可輕易的經由 L_p^* 得到。

2. 包含部份對照組資料的應用

由前一節可得知病例研究無法對基因與環境的主效應(main effect)、基因與基因的交互作用和環境與環境的交互作用等做統計推論。然而，當我們有對照組的部份訊息時，我們可以經由擴充我們的方法去估計主效應或是剩下的交互作用。我們的基本結論如下：假設條件獨立的性質成立。除了病例組的資料，如果我們有對照組中(E, W)((G, W))的觀測值，則我們可以利用最大概似法對 E(G)的主效應和 E×E、E×G、E×W(G×G、G×E、G×W)和 G×W×E 的交互作用做統計推論，其中(E, W)和(G, W)的資料來源可不相同。和前一節一樣，每個參數的統計推論可經由相對應的輪廓概似函數得到。

在基因與環境因子在對照組中條件獨立的假設下，假設我們額外有獨立的對照組觀測值 $\{(e_j^0,w_j^0),j=1,...,n_0\}$ 或 $\{(g_j^*,w_j^*),j=1,...,n_0^*\}$ 。當 n_0 = 0 但 $n_0^*\neq 0$ 時，我們可以證明輪廓概似函數爲

$$L_{pc}^*(G,W)=L_p^*\cdot\prod_{j=1}^{n_0^*}\prod_{m=0}^{M}\{f_0(G=g_j^*\,|\,W_m=1)\}^{W_{jm}^*}$$
$$\equiv L_p^*\cdot L_c^*(G,W)$$

此時，將 $\beta^*(t)$ 和 $b_m^*(t)$ 的表示式代入輪廓概似函數中，我們可以估計 β_k 、ϕ_{kr} 、θ_{kl} 、b_{km} 和 d_{klm} 。

二、修正基因型判定錯誤之病例分析

1. 機率模型與概似函數

在此我們假設基因因子 G，環境因子 E 與類別變數 W 都只有一個，且 G = g 表示樣本的眞實基因型包含 g 個高風險對偶基因(high risk allele)，g = 0, 1, 2。此外，與前一節相同，我們假設基因與環境因子在對照組中條件獨立。令 G_1^0 爲第一次的基因型觀測值，G_2^0 爲第二次的基因型觀測值。在病例研究中，如果樣本的基因型判定兩次，則我們可觀察到 E 和 (G_1^0,G_2^0)，如果只有判定一次，則我們觀察到 E 和 G_1^0 。

令 $f_d(w, e, g) = P(W = w, E = e, G = g\,|\,D = d)$, d = 0, 1，則我們假設一邏輯斯模型爲

$$f_1(w,e,g)=f_0(w,e,g)\exp(\alpha^*+\beta_g+\gamma e+\eta w+\theta_g e+\delta_g w+\xi ew).$$

其中 $\beta_0=\theta_0=\delta_0=0$ 。依據條件獨立的假設，我們得到

$$f_1(w,e,g)=f_1(w,e,0)\{f_0(+,g\,|\,w)/f_0(+,0\,|\,w)\}\exp(\beta_g+\theta_g e+\delta_g c),$$

其中 $f_0(+,g\,|\,w)=P(G=g\,|\,W=w,D=0)$ 。假設我們改寫式子

$$\{f_0(+,g\,|\,w)/f_0(+,0\,|\,w)\}=\{\mathrm{OR}_g\}^w R_g,$$

其中

$$\mathrm{OR}_g=f_0(+,g\,|\,1)f_0(+,0\,|\,0)/f_0(+,0\,|\,1)f_0(+,g\,|\,0),$$

$$R_g=f_0(+,g\,|\,0)/f_0(+,0\,|\,0).$$

則我們可得到

$$f_1(w,e,g)=f_1(w,e,0)\exp(\beta_g^*+\theta_g e+\delta_g^* w)$$

且在病例組中給定(W, E)，G 的條件機率爲

$$f_1(g\,|\,w,e)=\exp(\beta_g^*+\theta_g e+\delta_g^* w)\Big/\sum_{g'=0}^{2}\exp(\beta_{g'}^*+\theta_{g'}e+\delta_{g'}^* w),$$

其中 $\beta_g^*=\beta_g+\log(R_g)$ ， $\delta_g^*=\delta_g+\log(\mathrm{OR}_g)$ ，$g=1,2$，且 $\beta_0^*=\delta_0^*=0$ 。則我們可利用病例組的資料(W, E, G)和配適一多項式邏輯斯迴歸而得

到交互作用θ_g的統計推論。

接著我們假設在病例組中，基因型判定錯誤的機率(misclassification probability)爲 $P_1(G^0 = g^0 \mid G = g, W = w, E = e) = \phi(g, g^0)$和$(w, e)$獨立，其中$G^0$代表觀察到的基因型而$G$代表眞正的基因型。我們再假設給定眞實的基因型和$(w, e)$，第一次觀察到的基因型$G_1^0$和第二次觀察到的基因型$G_2^0$條件獨立，則在病例組中給定$W$和$E$，$G_1^0$的條件機率爲

$$P_1(G_1^0 = g_1^0 \mid W = w, E = e) = \sum_{g=0}^{2} \phi(g, g_1^0) f_1(g \mid w, e)$$

且給定W和E，(G_1^0, G_2^0)的條件機率爲

$$P_1((G_1^0, G_2^0) = (g_1^0, g_2^0) \mid W = w, E = e) = \sum_{g=0}^{2} \phi(g, g_1^0)\phi(g, g_2^0) f_1(g \mid w, e)$$

令Δ爲指標變數，$\Delta = 0$表示樣本的基因型只有判定一次則我們觀察到G_1^0，$\Delta = 1$表示樣本的基因型判定兩次則我們觀察到(G_1^0, G_2^0)。因此，我們利用資料$(\Delta_i, g_{1i}^0, \Delta_i g_{2i}^0, w_i, e_i)$，$i = 1, ..., n$，可建立一個條件概似函數爲

$$L = \prod_{i=1}^{n} P_1^{(1-\Delta_i)}(G_1^0 = g_{1i}^0 \mid W = w_i, E = e_i) P_1^{\Delta_i}((G_1^0, G_2^0) = (g_{1i}^0, g_{2i}^0) \mid W = w_i, E = e_i).$$

在此我們額外假設

$$P_1(\Delta = 1 \mid G_1^0, G_2^0, W, E) = P_1(\Delta = 1 \mid W, E),$$

這表示基因型判定兩次的機率只和(W, E)的值有關，且此機率對於參數的估計並無提供訊息。在大部份的應用中，機率$P_1(\Delta = 1 \mid W, E)$是一個常數。

上述的概似函數有一個重要的條件是兩次的基因型判定互相獨立且有相同的判定錯誤機率，Rice & Holmans [6]也假設此條件成立。然而有時兩次的基因型判定有相關性，此時在病例組中給定(W, E)，(G_1^0, G_2^0)的條件機率爲

$$P_1((G_1^0, G_2^0) = (g_1^0, g_2^0) \mid W = w, E = e) = \sum_{g=0}^{2} \pi(g_1^0, g_2^0 \mid g) f_1(g \mid w, e),$$

其中$\pi(g_1^0, g_2^0 \mid g) = P_1\{(G_1^0, G_2^0) = (g_1^0, g_2^0) \mid G = g\}$爲兩次基因型判定錯誤的聯合機率。除此之外，$G_1^0$給定$(W, E)$的條件機率與觀察的資料均和上述的相同，則我們可依照上述的步驟建構一個類似的條件概似函數再對交互作用做統計推論。

2. 概似比檢定與最大概似估計量

我們有興趣的參數是基因與環境因子的交互作用θ_g。依照概似函數與最大概似法，我們可得到概似比檢定統計量與最大概似估計量，其中最大概似估計量是必備的。參數的最大概似估計值可直接對概似函數取最大值而得到，但是在我們的概似函數中的參數數量相當多，因此我們建議使用 EM 演算法(expectation maximization algorithm)來求得參數的最大概似估計值。定義$N(g_1^0, g_2^0, g \mid w, e)$爲病例組中當$(W, E) = (w, e)$時，觀察的基因型$(G_1^0 = g_1^0, G_2^0 = g_2^0)$和眞實的基因型$G = g$的樣本個數。則多項式的概似函數爲

$$L^C = \prod_{w,e} \prod_{g, g_1^0, g_2^0} \left\{\phi(g, g_1^0)\phi(g, g_2^0)\right\}^{N(g_1^0, g_2^0, g \mid w, e)} \times \prod_{w,e} \prod_{g, g_1^0, g_2^0} \left\{f_1(g \mid w, e)\right\}^{N(g_1^0, g_2^0, g \mid w, e)}.$$

依照 Morris & Kaplan [7]的結果，上述的概似函數可簡化爲

$$L^C = \prod_{w,e} \prod_{g, g^0} \left\{\phi(g, g^0)\right\}^{K(g, g^0 \mid w, e)} \times \prod_{w,e} \prod_{g} \left\{f_1(g \mid w, e)\right\}^{N(g \mid w, e)}$$

其中

$$N(g \mid w, e) = \sum_{g_1^0=0}^{2} \sum_{g_2^0=0}^{2} N(g_1^0, g_2^0, g \mid w, e),$$

$$K(g, g^0 \mid w, e) = \sum_{g^*=0}^{2} \sum_{g_1^0=0}^{2} \sum_{g_2^0=0}^{2} N(g_1^0, g_2^0, g^* \mid w, e) \delta_{g, g^0}(g_1^0, g_2^0, g^* \mid w, e)$$

而$\delta_{g, g^0}(g_1^0, g_2^0, g^* \mid w, e)$表示當$(W, E) = (w, e)$時，眞實的基因型$G = g^*$與兩個觀察的基因型$(G_1^0, G_2^0) = (g_1^0, g_2^0)$中由眞正的基因型$g$判定爲$g^0$的個數。$\delta_{g, g^0}(g_1^0, g_2^0, g^* \mid w, e)$的值爲$\{0, 1, 2\}$的其中一個。以下將介紹 EM 演算法中的 E 步驟和 M 步驟。

E 步驟：

令 $\psi = (\beta_1^*, \beta_2^*, \theta_1, \theta_2, \delta_1^*, \delta_2^*)$ ，

且 $N(g_1^0 \mid w, e)$

$= \sum_{i=1}^{n} (1-\Delta_i) I(g_{1i}^0 = g_1^0, w_i = w, e_i = e)$ ，

和 $N(g_1^0, g_2^0 \mid w, e)$

$= \sum_{i=1}^{n} \Delta_i I(g_{1i}^0 = g_1^0, g_{2i}^0 = g_2^0, w_i = w, e_i = e)$

爲在病例組中的觀察值。給定 $\hat{\psi}^{(t)}$ 和 $\hat{\phi}^{(t)}(g, g^0)$ 爲第 t 步疊代得到的最大概似估計值，則第$(t+1)$步疊代中的期望資料爲

$$N^{(t+1)}(g_1^0, g_2^0, g \mid w, e)$$

$$= \frac{\hat{\phi}^{(t)}(g, g_1^0)\hat{\phi}^{(t)}(g, g_2^0) f_1(g \mid w, e, \hat{\psi}^{(t)})}{\sum_{g=0}^{2} \hat{\phi}^{(t)}(g, g_1^0)\hat{\phi}^{(t)}(g, g_2^0) f_1(g \mid w, e, \hat{\psi}^{(t)})} N(g_1^0, g_2^0 \mid w, e)$$

$$+ \frac{\hat{\phi}^{(t)}(g, g_1^0)\hat{\phi}^{(t)}(g, g_2^0) f_1(g \mid w, e, \hat{\psi}^{(t)})}{\sum_{g=0}^{2} \sum_{g_2^0=0}^{2} \hat{\phi}^{(t)}(g, g_1^0)\hat{\phi}^{(t)}(g, g_2^0) f_1(g \mid w, e, \hat{\psi}^{(t)})} N(g_1^0 \mid w, e)$$

M 步驟：

在第$(t+1)$步疊代中，$\phi(g, g^0)$ 的最大概似估計值爲

$$\hat{\phi}^{(t+1)}(g, g^0) = \frac{K_{g,g^0}^{(t+1)}}{\sum_{g^0=0}^{2} K_{g,g^0}^{(t+1)}},$$

其中

$$K_{g,g^0}^{(t+1)} = \sum_{w=0}^{1} \sum_{e=0}^{1} K^{(t+1)}(g, g^0 \mid w, e)$$

而 $K^{(t+1)}(g, g^0 \mid w, e)$ 和 $K(g, g^0 \mid w, e)$ 相同，只差在將式子中的 $N(g_1^0, g_2^0, g^* \mid w, e)$ 換成 $N^{(t+1)}(g_1^0, g_2^0, g^* \mid w, e)$。在此步疊代中，參數$\psi$的最大概似估計值 $\hat{\psi}^{(t+1)}$ 可由下列的概似函數得到

$$L^{(t+1)}(\psi) = \prod_{w,e} \prod_{g} \left\{ \frac{\exp(\beta_g^* + \theta_g e + \delta_g^* w)}{1 + \sum_{g^*=1}^{2} \exp(\beta_{g^*}^* + \theta_{g^*} e + \delta_{g^*}^* w)} \right\}^{N^{(t+1)}(g \mid w, e)}$$

其中 $\beta_0^* = \theta_0 = \delta_0^* = 0$ 而 $N^{(t+1)}(g \mid w, e)$ 和 $N(g \mid w, e)$ 相同，只差在將式子中的 $N(g_1^0, g_2^0, g \mid w, e)$ 換成 $N^{(t+1)}(g_1^0, g_2^0, g \mid w, e)$ 。

同樣因爲模型中的參數太多，我們以拔靴法(bootstrap procedure, Efron, & Tibshirani [8])來估計變異數的估計值。首先我們定義 $N(w, e)$爲觀察資料中$(W, E) = (w, e)$的個數。接著對每個$(W, E) = (w, e)$，拔靴法由 $P_1(\Delta = 1 \mid W, E)$ 產生 $N(w, e)$個指標資料Δ_i 且由

$$P_1((G_1^0, G_2^0) = (g_1^0, g_2^0) \mid W = w, E = e)$$

產生 $N(w, e)$個觀察的基因資料 (G_{1i}^0, G_{2i}^0)，其中產生資料所需要的參數值爲參數的最大概似估計值。重複產生 B 組資料我們可以得到 B 個θ_g的估計值 $\hat{\theta}_g^{(b)}$ ，則我們以 $\hat{\theta}_g^{(b)}$ 的經驗變異數(empirical variance)來估計 θ_g 的最大概似估計值的變異數，以 $\hat{\theta}_g^{(b)}$ 的百分位數(percentile)來估計信賴區間(confidence interval)。

上述的所有結果以及相關的討論請參考 Cheng [9,10]、Cheng & Chen [11]、Cheng & Lin [12]。

參考文獻

[1] W. W. Piegorsch, C. R. Weinberg and J. A. Taylor, *Stat Med*., **13**, 153 (1994).

[2] C. B. Begg and Z. F. Zhang, *Cancer Epidemiol, Biomarkers Prev*., **3**, 173 (1994).

[3] D. M. Umbach and C. R. Weinberg, *Stat Med*., **16**, 1731 (1997).

[4] P. S. Albert, D. Ratnasinghe, J. Tangrea and S. Wacholder, *Am J Epidemiol*, **154**, 687 (2001).

[5] B. G. Armstrong, *Epidemiol*, **14**, 467 (2003).

[6] K. M. Rice and P. Holmans, *Ann Hum Genet*., **67**, 165 (2003).

[7] R. W. Morris and N. L. Kaplan, *Genet Epidemiol*, **26**, 142 (2004).

[8] B. Efron and R. J. Tibshirani, *An introduction to the Bootstrap*, Chapman & Hall (1993).

[9] K. F. Cheng, *Stat Med*., **25**, 3093 (2006).

[10] K. F. Cheng, *Ann Hum Genet*., **71**, 238 (2007).

[11] K. F. Cheng and J. H. Chen, *Hum Hered*., **64**, 114 (2007).

[12] K. F. Cheng and W. J. Lin, *Stat Med*, **24**., 3289 (2005).