非干擾性質的風險因子

鄭光甫 主任/講座教授

干擾因子的考量起因於簡單的分析邏輯;假設我們使用統計迴歸分析方法去探討風險因子 x_1 對疾病結果 y 的影響效應時,我們必須確認影響效應是因為曝露於這個研究的風險因子而不是因為曝露於其他沒考慮的風險因子。因此,應用迴歸模型探討因子 x_1 對疾病結果 y 的影響效應時,假如存有第三個變數 x_2 也是疾病結果 y 的風險因子時,我們通常都會面臨是否要要將 x_2 放入迴歸模型共同分析的困難抉擇。若是放 x_2 進入迴歸模型的前後,發現 x_1 對疾病結果 y 的影響效應的估計有明顯的差異,則我們稱變數 x_2 是 x_1 及 y 的干擾因子(confounding factor)。否則,我們稱變數 x_2 是非干擾性質的因子(non-confounding factor)。若存在干擾因子,我們分析 x_1-y 關聯時必須用干擾因子作調整,否則分析會產生錯誤的結論。

很多人以為風險因子 x_2 是否為干擾因子和因子 x_1-x_2 間是否存在有關聯(非獨立)有莫大的干係。例如,在醫學的臨床實驗中,若是 x_1=1 或 0 分別代表治療組或控制組,由於隨機分派治療組或控制組的作法使得任何風險因子 x_2 和 x_1 顯得**獨立無關**;因為 x_1-x_2 間獨立互不存在影響,很多人就認定風險因子 x_2 是非干擾性質的風險因子。事實上,這種論點有部分是正確的有部分不是正確的。理論證明,若是風險因子 x_2 和 x_1 獨立無關的話,且使用的迴歸模型是**線性迴歸模型**,則 x_2 是非干擾性質的風險因子;若是使用的迴歸模型是**邏輯斯迴歸模型**,則 x_2 仍有可能是干擾的風險因子。

下面是一個研究抽菸 x_1 對肺炎 y 影響的案例,年齡 x_2 是肺炎 y 的風險因子(勝算比為 7.86),年齡-抽菸(x_1-x_2)的勝算比為 1,顯示 x_1 和 x_2 不存在關聯。邏輯斯迴歸模型中只使用抽菸 x_1 分析對肺炎 y 的影響效應時發現 x_1-y 勝算比為 7/3,但若同時使用 x_1 和 x_2 分析對肺炎 y 的影響效應時則發現 x_1-y 勝算比提高為 9/3。這個例子指出,在邏輯斯迴歸分析中即使 x_1 和 x_2 不存在關聯(獨立),x_2 仍然有可能是干擾因子。但是,理論也證明,在邏輯斯迴歸分析中若分別在 y=0 (非肺炎的族群)及 y=1 (肺炎的族群)下 x_1 和 x_2 都不存在關聯(即條件獨立)時,則 x_2 一定是

非干擾性質的風險因子。

	高齡	
	吸菸	非吸菸
肺炎	90	75
非肺炎	10	25

	低齡	
	吸菸	非吸菸
肺炎	50	25
非肺炎	50	75

	高齡	低齡
吸菸	100	100
非吸菸	100	100

	吸菸	非吸菸
肺炎	140	100
非肺炎	60	100

認定了 x_2 是一個非干擾性質的風險因子後，x_2 是否應該放在分析的迴歸模型和 x_1 一同研究？通常的答案是應該放，因為這樣做會使得迴歸模型的"合適性(goodness of fit)"更好，畢竟 x_2 是一個風險因子。但是，若我們研究的主要重點是在探討 x_1 對 y 影響的效應時(例如醫學的臨床實驗)，檢定效應是否存在？或效應的估計有多少？就是我們要分析回答的問題，模型是否合適不是最重要。此時，我們必須問的應該是：放 x_2 在分析的迴歸模型裡是否會加強檢定方法的檢定力？或降低估計方法的誤差？

以下我們分二種迴歸模型來討論不同的代表性做法。

線性迴歸模型的情況：

假設下面的二種線性迴歸模型，且 x_2 是一個非干擾性質的風險因子：

模型一， $\mu = E(y) = \beta_0^* + \beta_1^* x_1$ ， $\text{var}(y) = \sigma_1^2$ ；

模型二， $\mu = E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ ， $\text{var}(y) = \sigma_{12}^2$ 。

y 的期望值是 $E(y)$ ，變異數是 $\text{var}(y)$ 。傳統上，我們用最小平方法(y 為常態分配時即為最大概似估計法)估計 β_1^* 及 β_1 ，估計量記為 $\hat{\beta}_1^*$ 及 $\hat{\beta}_1$ 。由於模型二是 "正確"的模型， $E(\hat{\beta}_1^*) = \beta_1 + \beta_2 E(\sum(x_{1i} - \bar{x}_1)x_{2i})/(\sum(x_{1i} - \bar{x}_1)^2)$ ，又因為 x_2 是一個非干擾性質的風險因子($\beta_2 \neq 0$)，所以 x_2 和 x_1 必然是無任何的相關。惟， $\hat{\beta}_1^*$ 及 $\hat{\beta}_1$ 估計同樣的參數，但是他們的變異數不相同： $\frac{Var\hat{\beta}_1^*}{Var\hat{\beta}_1} = \frac{(1-\rho_{x_1,x_2}^2)}{(1-\rho_{y,x_2|x_1}^2)}$ ；

$\rho^2_{y,x_2|x_1}=\dfrac{\rho_{x_2,y}-\rho_{x_1,x_2}\rho_{x_1,y}}{\sqrt{1-\rho^2_{x_1,x_2}}\sqrt{1-\rho^2_{x_1,y}}}$，是在給定 x_1 下 x_2 和 y 的部分相關係數(partial correlation)，ρ_{x_1,x_2} 是 x_1 和 x_2 的 Pearson 相關係數。因為 x_2 和 x_1 無任何的關聯，所以 $\rho_{x_1,x_2}=0$，導致 $\dfrac{Var\hat{\beta}_1^*}{Var\hat{\beta}_1}=\dfrac{1}{(1-\rho^2_{y,x_2|x_1})}\geq 1$。這解釋為何**使用模型二有利**的原因(估計 $\beta_1(=\beta_1^*)$ 的誤差較小，檢定 $\beta_1(=\beta_1^*)=0$ 的檢定力較高)。

結論：若 x_2 是非干擾性質的因子($\beta_2\neq 0$)，探討 x_1-y 關聯的研究時使用模型二較好。

註：反過來，$\rho_{y,x_2|x_1}=0$(等同於 $\beta_2=0$)滿足時，$\dfrac{Var\hat{\beta}_1^*}{Var\hat{\beta}_1}=(1-\rho^2_{x_1,x_2})\leq 1$；表示，若是有 x_1 的模型中加入沒有解釋能力的因子時($\beta_2=0$ 表示 x_2 不是影響 y 的風險因子)，可能會導致 x_1 效應 $\beta_1(=\beta_1^*)$ 的估計誤差增大或檢定 $\beta_1(=\beta_1^*)=0$ 的檢定力下降。但若是 x_2 和 (x_1,y) 互相獨立的話則 $\rho_{x_1,x_2}=0$，和 $\rho_{y,x_2|x_1}=0$ 同時滿足，導致 $\dfrac{Var\hat{\beta}_1^*}{Var\hat{\beta}_1}=1$，因此 x_1-y 關聯的研究中使用模型一或二並無不同。

邏輯斯迴歸模型的情況下：

我們討論下面的二種邏輯斯迴歸模型；模型二是“正確”的模型：

模型一，　$\log\{\mu/(1-\mu)\}=\beta_0^*+\beta_1^* x_1$；

模型二，　$\log\{\mu/(1-\mu)\}=\beta_0+\beta_1 x_1+\beta_2 x_2$。

理論上我們可證明，在邏輯斯迴歸模型的情況下，假如 x_2 是非干擾性質的因子，則下面的條件之一會滿足或同步會滿足：(1)給定 y 時，x_1 和 x_2 獨立無關；(2)給定 x_1 時，y 和 x_2 獨立無關(等同於 $\beta_2=0$)。通常我們用最大概似估計法估計 β_1^* 及 β_1，估計量記為 $\hat{\beta}_1^*$ 及 $\hat{\beta}_1$。理論結果指出，若是僅有條件(1)滿足的話，則 $\dfrac{Var\hat{\beta}_1^*}{Var\hat{\beta}_1}<1$ 會成立，顯示模型二的作法會增加對 $\beta_1^*(=\beta_1)$ 估計的誤差，並且降低檢定 $\beta_1^*=0(=\beta_1)$ 的檢定力。**這個結果和線性迴歸的結果相反**。

結論：x_2 是非干擾性質的風險因子的話，在邏輯斯迴歸模型的情況下使用模型一

較好。

註：若是僅有條件(2)滿足的話($\beta_2 = 0$)，則 $\frac{Var\hat{\beta}_1^*}{Var\hat{\beta}_1} < 1$ 也會成立，即放入**無效應**的非干擾性質因子在邏輯斯迴歸分析中，對 $\beta_1^* (= \beta_1)$ 的估計誤差會增加，並且降低檢定 $\beta_1^* = 0 (= \beta_1)$ 的檢定力。

請特別注意，條件(1)和線性迴歸模型假設的："x_1 和 x_2 獨立無關"的條件是不同的。

最後，(1)和(2)同步滿足的話，則條件等同於"x_2 和 (x_1, y) 互相獨立"的條件，此時可證明 $\frac{Var\hat{\beta}_1^*}{Var\hat{\beta}_1} = 1$，即放或不放 x_2 在邏輯斯迴歸模型中均不會改變 $\beta_1^* (= \beta_1)$ 估計的誤差。

前面針對 x_1 - y 關聯的研究，討論是否要放非干擾性質的風險因子 x_2 進入迴歸模型中共同分析的優缺點。Neuhause 等人則特別額外考量 x_1 和 x_2 是否相關的情形，強調在臨床實驗研究(x_1 和 x_2 獨立無關)時，x_1 - y 關聯迴歸(廣義線性迴歸)分析中放入非干擾性質的風險因子 x_2 作分析可以提高檢定力以及降低估計誤差；另外，在世代研究時(x_1 和 x_2 相關)，非干擾性質的風險因子 x_2 不應放入迴歸模型中分析。

在其他的研究方法中是否使用模型一或二較有利？結論是和取得分析資料的抽樣方法有關：例如，邏輯斯迴歸模型的情況下，若是資料是病例-對照研究的資料，x_1 和 x_2 獨立無關，且疾病盛行率高(>20%)時，則使用模型二有較高的檢定力，但疾病盛行率很低僅有些許百分比時，則使用模型一經常有較高的檢定力。

參考資料

1. McCullagh, Peter and Nelder, John. (1989). *Generalized Linear Models, Second Edition*. Boca Raton: Chapman and Hall/CRC.
2. Henrik Madsen and Poul Thyregod. (2011). *Introduction to General and Generalized Linear Models*. Chapman & Hall/CRC.

3. Dobson, AJ and Barnett, AG. (2008). *Introduction to Generalized Linear Models* (3rd ed). Boca Raton, FL: Chapman and Hall/CRC.
4. Hardin, James and Hilbe, Joseph. (2007). *Generalized Linear Models and Extensions* (2nd ed). College Station: Stata Press.
5. Robinson, LD and Jewell, NP. (1991). Some Surprising Results about Covariate Adjustment in Logistic Regression Model. International Statistical Review, 59, 227-240.
6. Neuhause, MJ. (1998).Estimation Efficiency with Omitted Covariates in Generalized Linear Models. J American Statistical Association , 93, 1124-1129.
7. Pirinen, M, Donnelly, P and Spencer, CCA. (2012). Including Known Covariates can Reduce Power to Detect Genetic Effects in Case-Control Studies. Nature Genetics, 44, 848-851.