[研究新領域報導]

在伴隨變數有測量誤差下使用生物標記資料做工具變數的統計方法

美國弗雷得哈金森癌症研究中心公共衛生科學部　王清雲
逢甲大學統計學系　李燊銘

摘　要

在許多的醫學或流行病學的研究上，伴隨變數經常存在測量之誤差。此種資料可能引起在推論上正確性或解釋的困難性。當我們對研究個體的一個曝露變數(exposure variable)做測量，若無法藉由一個正規標準的測量來精準獲得，最常用的處理是使用重覆測量的輔助變數(surrogate variables)來取代。但在一些實際的例子中，對每一個研究的個體僅有一個輔助變數。探討此問題的確定關鍵，一般的處理方式，是尋找一個與觀測不到的曝露變數有相關的變數，其能有效的提供觀測不到曝露變數的訊息，此變數通常稱爲工具變數(instrumental variable)。在本文中，我們將介紹使用工具變數的一些重要應用，以調整有測量誤差存在之伴隨變數。我們也將藉由卜瓦松迴歸(Poisson regression)來描述處理的方法，及提供在此主題下一些新的研究議題。

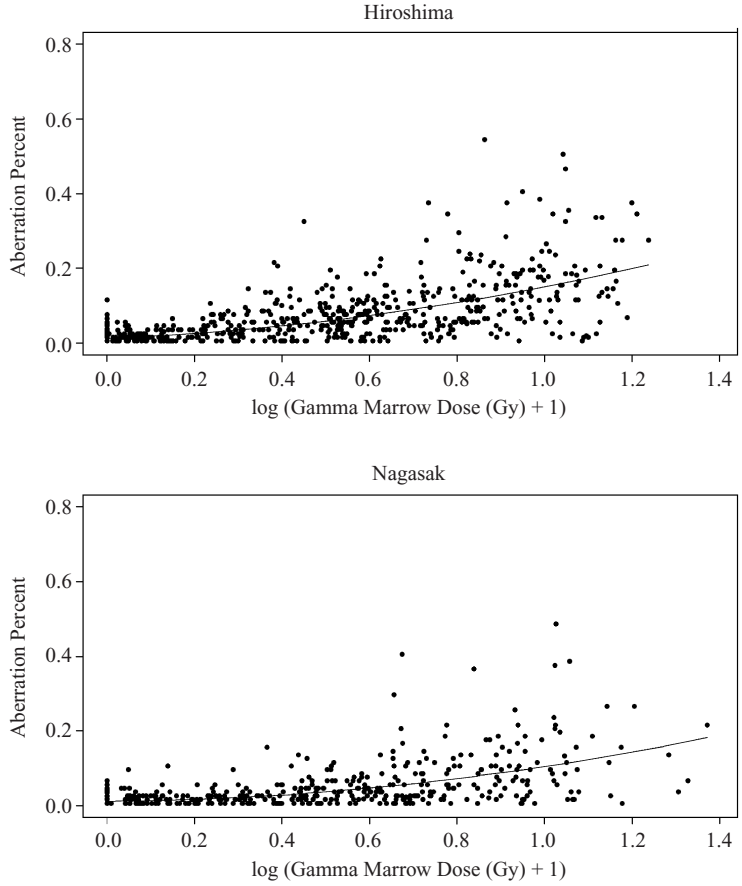
一、前　言

在流行病學的觀測研究，經常需面對的問題是要估計曝露變數與疾病的關係，然而實務上，經常遭遇到曝露變數是無法精確的獲得，這些問題有共同之處爲曝露變數是定量變數，且必需從個體曝露狀況的特徵中被測量或被估計。這些問題中，最重要的實例是在放射線流行病學的領域：人類被曝露在有離子化的放射線環境上，而目標器官上的劑量僅能是估計的，且在某些重要的公共健康之環境狀況（如環境的曝露、無法追蹤之職業傷害事故），曝露劑量估計經常是不可靠的。基於使用有測量誤差存在或不可靠的曝露變數，一般的認知可能導致估計曝露與疾病關係的偏誤，因此消除此偏誤仍是被視爲最適當的方法。此種的問題沒有特定一種最佳之方法可應用在所有情況，因爲在實際的應用上，最佳的處理方法應與可用資料的性質及引起曝露誤差之假設性質有關。

對輻射曝露與癌症之重要研究是輻射效應研究基金會(Radiation Effects Research Foundation, RERF)提供的，在 1947 年 RERF 及原子彈爆炸受難者家屬所組的委員會已對在 1945 年原子彈在長崎與廣島爆炸時之生還者做追蹤調查。RERF 在生還者及其子女所組成約 200,000 位之母體中，已檢查輻射曝露與疾病發生，細胞與遺傳的傷害，及其他的因素之間的關聯。壽命研究(The Life Span Study, LSS)爲 RERF 之研究項目的中心議題，而且此研究廣被認爲是提供研究輻射以及疾病之關係的重要資源。成年人的健康研究(Adult Health Study)是由近 20,000 位 LSS 會員之子集組成，此 RERF 的追蹤研究已提供從 1958 年起每隔 2 年的臨床檢驗資料。由於其有大量及長時間的追蹤的資料，因此提供了關於輻射對健康影響的估計及作爲輻射防護標準的最重要資訊。

在 RERF 追蹤研究中，曝露與疾病關係的分析是對個體於原子彈爆炸時，遭受到輻射劑量之估計有密切的依賴度。然而此劑量當然是無法測得，但可從有用的資料來估計，如透過面訪生還者所在位置、在爆炸時之遮避物、及分別在輻射源的物理計算、對透過空氣和遮避物之材料傳輸的結果來獲得估計之輻射量。當今的劑量測定法系統，是在 2002 年被提出的工具，命名爲 DS02，其是替代之前版本(DS86)，並能對 LSS 約 80,000 的成員估計出輻射的劑量。

Pierce et al. [9, 10]已對 LSS 資料探討了誤差校正之技術。而原子彈爆炸時之生還者遭受輻射



圖一　長崎與廣島兩城市的染色體畸變的百分比與 γ 射線劑量散佈圖。圖中曲線是從配適二次函數所獲得之估計

之劑量估計，在大多數分析中對隨機放射量之誤差是藉由誤差模型進行校正，此誤差模型是假設隨機誤差有標準誤爲眞實劑量的 35%。隨機誤差發生的原因是假設原子彈爆炸時，生還者對所在位置及遮避狀況之記錄有誤差所導致的。前面提到測量的校正法是建立在稱爲「迴歸校正」之方法上，此方法是利用在給定可觀測到的伴隨變數下的條件期望值來替代眞實無法觀測到的劑量（相關的文獻參見 Carroll et al. [4]，第四章）。

由於幅射劑量之不可靠性，故藉由收集生物標記資料可做爲有幫助之工具變數，主要的原因是因爲生物標記資料與無法觀測到的曝露有很好的關聯性。Stram et al. [12]由被 RERF 在 1968 至 1985 年間收集的 1,703 個個體，提供 T-淋巴細胞中呈現穩定性染色體畸變的百分比(%SCA)與無法觀測到之放射曝露、年齡、性別、曝露的年齡、檢驗的時間及城市曝露之效應分析。在之後，Kodama et al. [8]也對 3,042 個個案進行染色體畸變作檢驗調查研究，兩者之研究均發現在原子彈爆炸下，生還者的穩定性染色體畸變百分比(%SCA)與幅射劑量的關係是顯著的，提供了 %SCA 是有用且爲放射線曝露之長期標誌之令人信服證據。圖一中，我們使用 Stram et al. [12]所分析穩定之染色體資料，藉由二次函數建模來分別探討在長崎與廣島之案例中，異常細胞的百分比與 gamma 射線之對數劑量的關係。從圖中可看出異常細胞的百分比與估計的 gamma 射線對數劑量確有二次函數關係，從圖中也可看出誤差的變異數爲觀測不到的放射劑量或估計劑量之遞增函數，此證明%SCA 當輔助的工具變數時有異質之誤差。除此之外，放射線測量的誤差項有異質變異也是分析中重要的環節。

在本文中，我們研究在伴隨變數存在測量的誤差之迴歸分析。在校正的樣本中，我們可以運用不偏輔助變數以及工具變數。但是在非校正樣本中，我們只有不偏輔助變數但無工具變數。在第 2 節中，將對我們的迴歸問題做描述，此處是以卜瓦松迴歸來陳述我們的問題。迴歸校正之處

理則在第 3 節。第 4 節中，我們在校正樣本中研究工具變數的估計，一些重要的研究方向將在第 4 節說明。

二、迴歸模型

假設欲研究的個體共有 n 個個體，分別標爲個體 i, $i = 1, 2, \ldots, n$，令 X_i 爲眞正對疾病結果之曝露變數。設 Y_i 爲反應變數，其爲疾病的結果，假設 Y_i 是服從平均數爲 $\exp\{\beta_0 + \beta_1 X_i + \beta_2^T \boldsymbol{Z}_i\}$ 的卜瓦松分配，此處 $\boldsymbol{Z}_i$ 爲無測量誤差存在的伴隨變數向量。主要探討的議題是在給定對任意研究之個體，其伴隨變數 X 是無法觀測到的情況下，迴歸係數之估計。對每一個個體，我們假設存在可觀測的輔助變數 W_i 其來自可加性之測量誤差模型：

$$W_i = X_i + U_i,$$

此處 U_i 爲測量誤差項，$\mathrm{E}(U_i \mid X_i) = 0$，及 $i = 1, \ldots, n$。爲了符號的簡便，此處我們僅考慮 X_i 爲一純量變數。在上述的模型中我們假設僅有一個 W_i 之值，但在後面提供的方法，也可應用在有重覆測量之情況。除了此不偏輔助變數 W_i 之外，我們假設在研究的子群中，尚有工具變數(instrumental variables)可使用，此子群稱爲校正樣本(calibration sample)。設工具變數以 M_i 表示，其滿足

$$M_i = h(X_i, \boldsymbol{Z}_i) + V_i,$$

此處 V_i 爲誤差項，$\mathrm{E}(V_i \mid X_i) = 0$ 及 $h(X, \boldsymbol{Z})$ 爲已知之函數但其含未知的參數，例如 $h(X_i, Z_i) = \alpha_0 + \alpha_1 X_1 + \boldsymbol{\alpha}_2^T \boldsymbol{Z}_i$。在本文中，我們假設在校正樣本中，僅有一個工具變數 M_i 之值可利用，但後面之研究均可推展到有重覆測得 M_i 之值的情況。在我們欲研究估計迴歸係數的方法之前，應先討論上述模型之確認性(identifiability)問題。

(1) 若有重覆之 W_{ij} 可運用，則參數是可確認的；此是一般在測量誤差模型之情況，此藉由重覆之測量估計個體內及個體間的變動來獲得確認性成立。
(2) 若 W_i 沒有重覆測量及對所有研究之個體中均無 M_i 之值可運用，僅有 Y_i 及 W_i 之訊息，則參數是不可確認的。
(3) 若僅有一次之 W_i 及在校正樣本中，工具變數 M_i 之值也僅有一個可運用時，則參數是可確認的，因此可結合 Y_i，W_i 及 M_i 以估計所有之參數。

上述有關確認性問題之討論，可參見 Carroll et al. [3]之相關研究。

三、迴歸校正

一個非常普通且自然的概念，是在分析時直接使用 W_i 來替代觀測不到的 X_i。在測量誤差存在時，直覺(naïve)估計量會導致嚴重偏誤是眾所知曉的結果，而估計量偏誤的評估，可由

$$\mathrm{E}(Y \mid W_i, \boldsymbol{Z}_i) = \mathrm{E}\{\exp(\beta_1 X_i) \mid W_i, \boldsymbol{Z}_i\} \exp(\beta_0 + \beta_2' \boldsymbol{Z}_i)$$

來獲得。上式中 $\mathrm{E}\{\exp(\beta_1 X_i) \mid W_i, \boldsymbol{Z}_i)\}$ 爲動差母函數(moment generating function)。對任意之隨機變數 X 及 Z，令 μ_x 爲 X 的平均數，Σ_{xz} 爲 X 與 Z 的共變異數，及 σ_x^2 爲 X 的變異數，當 $(X_i, U_i, \boldsymbol{Z}_i^T)$ 來自一多元常態分配，則可很容易獲得

$$\mathrm{E}(X \mid W, \boldsymbol{Z}) = \mu_x + (\Sigma_{xw}, \Sigma_{xz}) \begin{pmatrix} \sigma_w^2 & \Sigma_{wz} \\ \Sigma_{wz} & \sigma_z^2 \end{pmatrix}^{-1} \begin{pmatrix} W - \mu_x \\ \boldsymbol{Z} - \mu_z \end{pmatrix}.$$

因此，在上述的常態假設下，可看出直覺估計量是有偏誤的。此外，另一種常用的處理方法是利用 $\mathrm{E}(X_i \mid W_i, \boldsymbol{Z}_i)$ 替代 X_i，以做爲斜率估計之調整。在計算校正函數 $\mathrm{E}(X \mid W, \boldsymbol{Z})$ 時，將隱含一些干擾參數 μ_x, σ_x^2 及其他之干擾參數。若測量誤差項 U，其變異數爲 σ_u^2 是一常數時，這些干擾參數可用資料 $Y, W, \boldsymbol{Z}$ 及 M 的相關動差來估計。在 Wang, et al. [14]及 Carroll, et al. [4]（第四章）文獻中，提出以 X 的條件期望值 $E(X \mid W, Z)$ 的估計值來替代 X 之值，此稱爲迴歸插補校正估計法。有一些問題很自然地會被提出：

(1) 當常態的假設不正確時，其結果如何？
(2) 如何估計 β_0？
(3) 如何處理測量誤差有異質變異數(heteroscedastic measurement errors)之情況？

在下節中，我們所提供方法將可回應上述之三個問題。

四、校正樣本中有工具變數之估計

當輔助變數 W 是不偏且有工具變數可運用

下，Buzas and Stefanski [2]提出條件分數之估計方法。條件分數之估計是假設誤差項 U_i 及 V_i 的分配是常態分配，但對於伴隨變數則不需有分配上之假設。Buzas [1]對非線性的迴歸模型提出加權的工具變數估計量，他的主要觀念，是尋找一個純量值的加權函數，以獲得不偏的估計方程式。令 $\eta_i = 1$ 表示第 i 個個體是在校正樣本中，$\eta_i = 0$ 表第 i 個個體不在校正的樣本內，當 W 及 M 僅有一個測量值下，在校正樣本內，Buzas 所提出方法的估計方程式可寫成

$$\sum_i \eta_i \omega(W_i, M_i, \boldsymbol{Z}_i) \begin{pmatrix} 1 \\ M_i \\ \boldsymbol{Z}_i \end{pmatrix} \{Y_i - \exp(\beta_0 + \beta_1 W_i + \beta_2^T \boldsymbol{Z}_i)\} = 0,$$

此處

$$\omega(W_i, M_i, \boldsymbol{Z}_i) = \exp\{0.5\beta_1 \mathrm{E}(W_i \mid M_i, \boldsymbol{Z}_i) - 0.5\beta_1 W_i\}.$$

Buzas [1]證明：只要測量誤差項 U_i 的動差母函數存在且 U_i 的分配對稱於 0，上述 β 的估計是具有一致性的。當給定(M_i, Z_i)下，W_i 是線性的，則可用在給定(M, Z)下 W 的迴歸式以求出 $E(W_i \mid M_i, Z_i)$ 之估計，此種估計稱爲加權工具變數估計量(weighted instrumental variable (WIV) estimator)。

在許多實務上的問題中，輔助變數之測量誤差項 U_i 及工具變數之測量誤差項 V_i 具有異質的變異數。例如測量誤差項的變異數爲眞正觀測不到之曝露變數的遞增函數，在圖 1 中，在給定放射劑量下異常細胞百分比之迴歸及工具變數之變異數均是眞實放射劑量的遞增函數。在伴隨變數有測量誤差問題中，若測量誤差項的變異數不是固定常數時，如何提供一合宜的估計方法是一個重要的研究課題。

五、重要研究課題之討論

本文中在卜瓦松迴歸下，當校正樣本中有工具變數可運用時，我們回顧一些有測量誤差存在時之估計方法。有一些重要可能的研究課題，可做爲未來研究之方向：

(1) 在邏輯斯迴歸中，伴隨變數有測量誤差時且有工具變數可運用下，泛函法(functional methods)、無母數法：Huang and Wang [6, 7]在相關的問題中，已提出無母數法來校正參數的估計，但他們並未考慮工具變數的使用，因此未來是值得研究之方向。

(2) 在 Cox 迴歸中，伴隨變數有測量誤差時且有工具變數可運用下，泛函法、無母數法：Huang and Wang [6, 7]在相關的問題中，已提出無母數法來校正參數的估計，但他們並未考慮工具變數的使用，此在存活分析中是很重要的問題。

(3) 在(1)及(2)的主題中，當測量誤差或工具變數之誤差可能具有異質性時之無母數法，即當無法觀測到之曝露變數，其存在測量或工具變數之誤差項與其他變數有關時，無母數法之估計問題。

(4) 對存在的估計方法應用在藉由飲食攝取之生物標記資料，探討飲食攝取與疾病之關聯。飲食評量量表(food frequency questionnaire)，飲食記錄(food records)，或 24 小時之回顧爲傳統之飲食攝取之資料收集方法，而對生物標記應用在飲食攝取之資訊，是對觀測不到的眞實之攝取量提供客觀的輔助工具。例如，雙標記水(doubly labeled water, DLW)做爲能量消耗的生物標記及尿中的氨量當做蛋白質之消耗之生物標記。在此主題之相關文獻，可參見 Wang [13]。

(5) 對存在的估計方法應用在由運動量之生物標計資計，探討運動量與疾病之關聯。問卷是最常被用來測量運動量之方法，但使用生物標記資料做爲無法觀測到之運動量之輔助工具是很重要的研究議題。

致　謝

此研究是由美國國家衛生研究院(NIH)及美國國家科學院(NAS)所支持王清雲研究員之經費，計劃編號 CA 53996，及 DELS 749706，另外也由台灣之國家科學委員會數學推動中心支助王清雲研究員來台灣訪問之經費。

參考文獻

[1] J. S. Buzas, *Communications in Statistics, Ser. A*, **26**, 2861 (1997).

[2] J. S. Buzas and L. A. Stefanski, *Amer. Statist. Assoc.*, **91**, 999 (1996).

[3] R. J. Carroll, D. Ruppert, C. M. Crainiceanu,

T. D. Tosteson and M. R. Karagas, *J. Amer. Statist, Assoc*., **99**, 736 (2004).

[4] R. J. Carroll, D. Ruppert, L. A. Stefanski and C. M. Crainiceanu, *Nonlinear Measurement Error Models, Amodern Prespective*, second edition, Chapman and Hall, London (2006).

[5] Y. Huang and C. Y. Wang, *J. Amer. Statist. Assoc*., 95, 1209 (2000).

[6] Y. Huang and C. Y. Wang, *J. Amer. Statist. Assoc*., **96**, 1469 (2001).

[7] Y. Huang and C. Y. Wang, *Statistica Sinica*, **16**, 861 (2006).

[8] Y. Kodama, D. Pawel, N. Nakamura, D. Preston, T. Honda, M. Itoh, M. Nakano, K. Ohtaki, S. Funamoto and A. A. Awa, *Radiation Research*, **156**, 337 (2001).

[9] D. A. Pierce, D. O. Stram and M. Vaeth, *Radiation Research*, **123**, 275 (1990)

[10] D. A. Pierce, D. O. Stram, M. Vaeth and D. Schafer, *J. Amer. Statist. Assoc*., **87**, 351 (1992).

[11] D. L. Prenston, Y. Shimizu, D. A. Pierce, A. Suyama and K. Mabuchi, *Radiation Research*, **160**, 381 (2003).

[12] D. O. Stram, R. Sposto, D. Preston, S. Abrahamson, T. Honda and A. A. Awa, *Radiaaation Research*, **36**, 29 (1993).

[13] C. Y. Wang, *Scandinavian Journal of Statistic*, **35**, 613 (2008).

[14] C. Y. Wang, L. Hsu, Z. D. Feng and R. L. Prentice, *Biometrics*, **53**, 131 (1997).

[15] M. Yamada, F. L. Wong, S. Fujiwara, M. Akahoshi and G. Suzuki, *Radiation Research*, **161**, 622 (2004).