

測度誤差模型簡介

中央研究院統計科學研究所 鄭紀倫

導論

統計中利用各種不同方式蒐集數據再進行分析，但在蒐集數據中常有所謂測量誤差 (measurement error) 產生。迴歸模型無疑的是統計中理論與實際上均是相當重要的模型，但迴歸模型中的自變數 (independent variable) 無論是固定 (fixed) 或是隨機 (stochastic) 均是已知，也就是說無測量誤差，如果以數學式表示 (為簡單起見，以單元迴歸來論)

$$y_i = b_0 + b_1 x_i + e_i \quad i = 1, \dots, n \quad (1)$$

其中 x_i 為自變數， y_i 為應變數 (dependent variable)，而 e_i 為誤差 (error)，且 x_i 與 e_i 是獨立的 (independent)。

測度誤差模型 (measurement error model) 則是假設我們觀察到 x_i 與 y_i ，但兩者均有測度誤差，所以

$$y_i = h_i + e_i, \quad x_i = x_i + d_i, \quad i = 1, \dots, n \quad (2)$$

其中 h_i 是真實的應變數， x_i 為真實的自變數， x_i 一般又稱為潛在變數 (latent variable)，而真實的應變數與真實的自變數的關係為

$$h_i = b_0 + b_1 x_i \quad i = 1, \dots, n \quad (3)$$

模型(2)-(3)一般稱為古典測度誤差模型 (classical measurement error model)。舉個最簡單的例子：

質量=密度×體積

質量(mass)及應變數，體積(volume)為自變數，密度即為“斜率” b_1 (本式中截距 (intercept) b_0 為 0)。很顯然地，在測量質量與體積中均會有測量誤差，所以用模型(2)-(3)是比(1)來的恰當一點。如果仔細分析測度誤差模型，我們可以以 x_i (latent variable) 的屬性分為兩大類函數：

- (i) 函數模型 (functional model): x_i 為固定且未知的參數，
 - (ii) 結構模型 (structural model): x_i 為隨機 (但為 IID)。
- 模型中 x_i 是種 nuisance 參數，我們對它本身不

感興趣。在估計中如 MLE (maximum likelihood estimate)，因 x_i 的數目隨樣本數 (n) 而變大至無窮，此為著名的 Neyman-Scott problem，會給 MLE 帶來很大困擾；在結構模型中 x_i 為 IID，乍看之下好像比較容易處理，比如說 (x_i, d_i, e_i) 為 IID 常態 (normal) 分佈，我們可以用 ML 來估計，不過仔細一看參數為 $m = EX_1, s^2 = VarX_1, s_d^2 = Var d_1, s_e^2 = Vare_1, b_0, b_1$ ，而常態分佈的 sufficient statistics 只有 $\bar{x}, \bar{y}, S_{xy}, S_{xx}, S_{yy}$ 等五個，直觀上來說，ML 估計會出問題，事實上也是如此。

估計

在詳述 ML 估計的問題之前先說一下所謂確認性 (identifiability)。簡單地說在一個參數模型 (parametric model) 中，模型被參數 q (向量) 決定 (例如在常態中 $q = (m, s^2)$)，如果有 $q_1 \neq q_2$ 但產生的 realization 模型 $P_{q_1} = P_{q_2}$ ，我們說此模型 (參數) 是不確認的 (unidentifiable)。這便是說有兩組的參數 q_1, q_2 但可產生相同的 realizations (即 data set)。參數如果不確認，會產生立即的困難。因我們觀察到 realizations $\tilde{z} = (z_1, \dots, z_n)$ (即樣本)，要從 $\tilde{z}$ 去估計 q (真實的參數)，無論用何種方式，MLE、method of moments、... 等等得到 $\hat{q} = \hat{q}(\tilde{z})$ ，那 $\hat{q}$ 是不可能一致的 (consistent)。因無論強一致 (strong consistent) 或若一致 (weak consistent) 收斂值應是唯一的。但 $q_1 \neq q_2$ 的情形下，根本不可能找到一致的估計值，所以一般說來確認性是一個參數模型的“最低需求”。

Reiersol [6] 證明了結構性模型得確認性如下：

“假設 (d_i, e_i) 是常態分佈，那 b_1 是確認的若且唯若 x_i 是非常態分配”。這個結果使得在常態分佈下的 (x_i, y_i) 沒有任何不能找到一致性的估計。一般來說，我們必須加上一些所謂的確認條件，常用的有

1. $l = s_e^2 / s_d^2$ 已知，或

2. s_d^2 已知, 或
3. s_e^2 已知, 或
4. s_d^2 和 s_e^2 均已知, 或
5. 截距 b_0 已知。

當上述任何一項條件成立則 MLE 即可算出, 這些條件我們只能從以往的經驗或者有重複的實驗, 比如

$$x_{ij} = x_j + d_{ij}, \quad j = 1, \dots, k, \quad i = 1, \dots, n$$

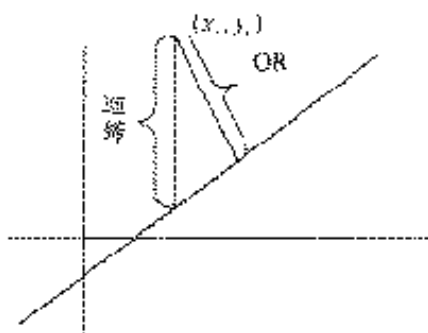
中可估計 s_d^2 , 如果是多重的測度誤差模型

$$y_i = b_0 + b_1' x_i + e_i, \quad x_i = x_i + d_i, \quad i = 1, \dots, n$$

x_i, x_i, d_i, b_i 為 $p \times 1$ 向量, 而 y_i, x_i 仍為一維 scalar, 則上述的確認條件將化為:

1. $\Omega = \text{cov}(d', e) = d' W_0, \quad d \neq 0, \quad W_0$ 已知, 或
2. S_{dd} 已知。

嚴格來說, 測度誤差模型早在一百多年前 (Adcock [1]) 就被提出, 不過是以幾何方式、傳統的迴歸模型是求



$$Q(b_0, b_1) = S(y_i - b_0 - b_1 x_i)^2$$

的最小值, 而“垂直迴歸”(orthogonal regression 簡稱 OR)是求

$$Q'(b_0, b_1) = S \left(\frac{y_i - b_0 - b_1 x_i}{\sqrt{1 + b_1^2}} \right)^2$$

的最小值。

當然, 上述只是幾何上的考量, 而 OR 所對應的統計模型即是測度誤差模型中假設 $s_e^2/s_d^2 = l$ 中 $l = 1$ 的情形。很多人認為 OR 即是測度誤差模型, 但此一觀念是不正確的, OR 只是此模型中的一小部份而已。

一百多年來 OR 被屢次“重新發現”, 一直到 50 年代才有對整個測度誤差比較完整的研究。到了 1980 年代, 線性的測度誤差模型的點估計

已被完整的研究, 包括常態分佈中的 MLE 及非常態的 method of moments 等。

在區間估計方面的進展很小, 尤其是 Gleser and Hwang [5] 發現一個很重要的問題, 大意是“不管如何取得區間估計, 總有不為零的機率存在, 使得 b_1 的區間估計長度為無限”。也就是說 b_1 的區間估計很可能為整個實數! 反過來說, 如果用 asymptotic 方式來做區間估計, 雖一定得到有限長度的區間但 confidence level 就可能不是預定得值(如 90 %), 此一問題嚴重的影響區間估計的準確性, 一般的人比較難了解此一問題的複雜性, 如果從直觀來看

$$x = x + d$$

x 在工程上叫訊號(signal), d 叫雜訊(noise)而 x 是觀測值, 直覺上訊號比雜訊來的小或差不多, 那觀測值相當不可靠, 在計量心理學(psychometrics)上叫 $s_x^2/s_x^2 \equiv t$ 的可靠度(reliability), 可靠度愈接近 1, 那問題就小, 如果低於 0.5, 那問題就變的大了, 這也是上述區間在估計出現問題的癥結。

自 1980 年代中晚期開始, 非線性誤差模型得到很大的關注, 尤其是生物醫學統計, 如廣義線性模型(GLM)用途很大。但實際運作中, 自變數 x 有測度誤差是很常見的, 一般傳統的 GLM 模型不適用, 而有測度誤差的 GLM 模型就被廣泛使用。

不過非線性測度誤差模型的難度很高, 雖然有許多論文發表, 但仍有很多問題仍待解決, 而且絕大多數的研究也只限於點估計而已。讀者可參閱 Carroll, Ruppert and Stefanski [2] 的專書 Measurement Error in Nonlinear Models。

其他相關研究

在社會科學有所謂的 factor analysis, 在計量經濟中的 simultaneous equation model 均是與測度誤差模型有關, 此外在數值分析中的 total least squares (TLS) 亦與此有關, 後者是 80 年代以來的新興議題。TLS 主要研究用於 signal process, control theory 等工程方面, 所以對 computation 方面特別注重, 統計方面的味道就較小了, 請見 Van Huffel and Vandenalle [9]。

在 Cheng and Van Ness [3]也對上述相關議題作了一簡介, 事實上上述任何一個議題都可說是一個小的“學門”, 要深入了解就會變成另一個研究領域了。

結語

測度誤差模型的出現是很自然的事，但此模型用意並非在取代傳統的迴歸模型，其真正作用是在使用一般迴歸模型時若所得結果似乎有問題，或者在蒐集數據時即發現測度誤差太大無法忽視，測度誤差模型在此情形下是一個重要的 alternative。

雖然自 80 年代晚期非線性測度誤差模型被廣泛的研究，但在線性模型中許多重要的議題如 diagnostics, variable selection 等等均欠缺研究。這是因為此類議題難度相當高，在目前仍未有理想的結果。這些問題與非線性誤差模型均是未來重要的研究方向，大家可參考 Stefanski [8] 所做的評論。

關於測度誤差模型的參考書有 Schneeweiss and Mittage [7], Fuller [4], Carroll, Ruppert and Stefanski [2] 和 Cheng and Van Ness [3]。其中入門書籍以 Fuller [4] 與 Cheng and Van Ness [3] 為主，前者內容較多但資料比較舊，後者是以方法論方式來寫，較少證明，而對觀念有較多的著墨。至於 Schneeweiss and Mittage [7] 是德文，計畫中的英文新版尚未出書。而 Carroll, Ruppert and Stefanski [2] 是以非線性模型為

主，對初學者較不宜，但對非線性誤差模型有興趣者是最佳選擇。

參考文獻

- [1] R. J. Adcock, *Analyst*, **4**, 183 (1877).
- [2] R. J. Carroll, D. Ruppert and L. A. Stefanski, *Measurement Error in Nonlinear Models*, Chapman & Hall, London (1995).
- [3] C. L. Cheng, and J. W. Van Ness *Statistical Regression with Measurement Error*, New York, Oxford University Press (1999).
- [4] W. A. Fuller, *Measurement Error Models*, Wiley, New York (1987).
- [5] L. J. Gleser and J. T. Hwang *Ann. Statist.*, **15**, 1351 (1987).
- [6] O. Reiersol *Econometrica*, **18**, 375 (1950).
- [7] H. Schneeweiss and H. J. Mittag *Lineare Modelle mit feherbehafteten Daten*, Physica-Verlag, Heidelberg (1986).
- [8] L. A. Stefanski *J. Amer. Statist. Assoc.*, **95**, 1353 (2000).
- [9] S. Van Huffel and J. Vandewalle *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia (1991).