知識天地

測度誤差模型簡介

鄭紀倫（本院統計科學研究所研究員）

導論

任何測量、資料蒐集甚至分析演算，均存在誤差，但這些誤差常被忽略。最可能的原因爲這些誤差均非已知，而無從列入。但如果誤差「太大」，那結果的分析與研判將會大受影響。廣義來看，測量誤差存在於各行各業，但往往爲人所忽略。

本文所介紹的是比較狹義的測量誤差模型（Measurement Error Model 或 Errors-in-Variables Model）。在統計理論與應用中，迴歸模型（Regression Model）扮演極重要的角色。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。用符號表示即爲應變數（Dependent Variable）Y，自變數 X 均爲已知的觀測值。迴歸模型即是假設 Y 和 X 之間有關係，最簡單的例子即是線性模型。而測量誤差模型是認爲觀測值是 Y 和 W，但 Y 和 W 之間的關係並不清楚，因 W 是 X 的替代值。亦即 X 才是真實值，但我們觀察不到，只能測到 X 的替代品 W。這也就是說明在測量 X 時是存在誤差 δ ，最簡單的情形是 $W = X + \delta$ 。而 (Y, W, X) 所構成的模型即爲測量誤差的模型。

最原始的線性測量誤差模型在 1870 年代即已出現，但並未受到特別重視。其後百餘年，迴歸模型無論在理論，應用上均有長足的進步，而且成爲統計分析中極爲重要的工具。反觀測量誤差模型卻進展緩慢，真正的原因很難釐清，但模型的複雜度可能是主因。直觀來說，我們能蒐集到的資料是以 Y、W 的形式出現，但 Y 和 W 之間的關係不明，我們只知道 Y 和 X 之間的關係。在此情形下，任何統計推論或多或少都會碰到難關。技術性上來說，從模型參數的估計（點估計和區間估計）均有一定的困難和障礙。這也是一般的應用上，測量誤差模型是不受重視。而一般統計軟體，也沒有此類的設計。到了上世紀 80 年代，測量誤差模型開始受到重視，其原因是在許多資料用迴歸模型處理時，結果十分不理想，究其原因是自變數中的誤差太大無法以傳統的迴歸模型來分析。

過去 20 多年來，有關測量誤差模型的研究如雨後春筍般興起。但嚴格來說，具突破性進展論文並不多，大多是針對已知結果做些微的改進。這也是因爲在許多學門中，如化學、工程、醫學等系中均須用測量誤差模型在處理問題，而加速了此問題的研究。模型愈複雜，研究的難度也大爲提高。不過也正因如此，測量誤差模型的研究也變爲極具挑戰性（Challenging）的問題。

此外，在此筆者要強調一點，測量誤差模型和在統計上另一個重要領域－資料缺漏（Missing Data）問題，不可混爲一談。Missing Data 一般是指在資料上有若干比例（如 5%）中 Missing，但我們要利用現存的不完整資料來分析，對缺漏部份要做一些補救。如果一定要說測量誤差模型和 missing data 的關聯，我們可以說在資料中所有的自變數 X 均爲 Missing。此 2 問題基本上是不同的領域，而且研究的方法也大爲不同。

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少了，請見 Van Huffel and Vandenalle [1]。

在 Cheng and Van Ness [2]也對上述相關議題作一簡介，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 20 世紀 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。這是因爲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 [3]所做的評論。

關於測度誤差模型的參考書有 Schneeweiss and Mittage [4],Fuller [5],Carroll et al. [6]和 Cheng and Van Ness [2]。其中入門書籍以 Fuller [3]與 Cheng and Van Ness [2]爲主，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 [4]是德文，計畫中的英文新版尚未出書。而 Carroll et al. [6]是以非線性模型爲主，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] Van Huffel, S. and Vandewalle, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis.* SIAM, Philadelphia.

[2] Cheng, C. L. and Van Ness, J. W. (1999). *Statistical Regression with Measurement Error*. New York, Oxford University Press.

[3] Stefanski. L. A. (2000). *Measurement Error Models.* J. Amer. Statist. Assoc. 95. 1353-1358.

[4] Schneeweiss, H. and Mittage, H. J. (1986). *Lineare Modelle mit feherbehafteten Daten.* Physica-Verlag, Heidelberg.

[5] Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.

[6] Carroll, R. J., Ruppert D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.