資料檢視與清理

針對抽樣及統計上容易發生的客觀性偏差,美國加州大學教授Tabachnick和Fidell在1983年發展出一套「資料清理（Data Cleaning up）」技術,以檢視和修正有問題的數據,穩定推算結論的精確程度。嚴格說起來,這套技術並非這兩位學者的創見,而是博採眾多學者思考的成果,整理組織成一套完整的程序,具有十分實用的效益。

本研究者在此將這項技術再進一步發展、詮釋、與介述應用的方法。

Tabachnick & Fide（1983）對資料清理建議的要旨及特色為:

（1）各種多變項分析幾乎全需借助相關矩陣,而矩陣中相關係數之值可能因為:

資料正確性; 迷失資料處理; 偏頗資料; 資料分配

是否符合統計假設;而產生高估或低估的現象,因此宜就以上四方面檢視資料,作一先期之評估。以檢視資料為手段,而達成校正矩陣、提高後續多變項分析精確程度的目的。

（2）強調系統化檢視。一般研究者在執行分析,雖然通常也會對原始資料及基本分配作一檢視,但經常出之粗放,亦不詳細報告結果,對「線性假設」「異質假設」等,常抱著先驗認可的態度。而 Tabachnick & Fidel則認為必需一絲不苟,步步逐項檢視,他們雖未提出驚人的新意,卻提供了系統化的檢視架構。

（3）提供完整檢視程序及方法。每一個檢視項目,均有學者發展出不同的檢視與處理方法,Tabachnick & fidel 縷列各家學說並比較優劣,說明其適用時機。

（4）強調謹慎態度;追求精確資料而不是操作資料。在檢視處理過程中,有不少學者發展出人為轉換資料的方式,可以使資料變得很「漂亮」,卻有損資料真實面貌,Tabachnick & Fidel雖介紹了這些方法,卻一再提示不宜輕易採用。他們有一句名言:「淨化資料很重要,但不是要改變它。」

一、檢視內容檢視工具與清理特異值

Tabachnick & Fidel 雖然詳細介紹了檢視和處理資料的內容,但檢視的工作很繁鉅,非人工能夠達成,必需借電腦的協助。

本文作者將國內易於取得的SPSS中可供檢視的程式或指令,與檢視內容並列於「表1」中,以明眉目。

表1.

檢視內容	檢視工具
1.資料正確性 （1）不合理值	Frequencies and / or Condescriptive

（2）類別分配 （3）樣本分配	
2.迷失資料處理	個別程式選項
3.雙變項偏頗資料處理	Regression 的 Scatter gram
4.非線性和異質性	同上
5.相依性和一向性	個別程式選項

經過以上檢視，如果發現有少數特異值(outlier)，就應在資料集中將其刪除，是為資料清理。

二、資料正確性

（1）不合理值

類別資料：是否無此類別水準，如「性別」出現「3」等。

連續資料：最大、最小值合理性。

（2）類別資料分配

類別細格內，若數值 <5，則此類別水準應考慮合併。

（3）連續資料樣本分配

是否太偏，而嚴重不符常態分配。當樣本數有限時，不易形成明顯的常態分配；所謂嚴重不符，包括成為「凹」形、或「』」形等。

三、迷失資料處理

迷失資料百分比是否太高？

在一般研究經驗中,八成左右的完整資料均視為可容忍,不必考慮過多人為操作。

迷失資料的後續處理法有三種 ：

1. 樣本全案刪除（Listwise Deletion）
2. 變項配對刪除（Pairwise Deletion, Tabachnick &Fidel 則稱為相關矩陣處理法Correlation Matrix Method）
3. 估計迷失值。

如果樣本數不算頂多,樣本全案刪除,損失很大,不宜採用。而估計迷失值的操作色彩較濃,本項研究完整樣本的比例尚高,不必冒操作風險。因此,多數時機採用「變項配對刪除法」,應屬最佳選擇。即僅在遇到某一變項含有迷失值時,才刪除這項變項,以保留資料的豐富性。

四、雙變項偏頗資料的處理

雙變項連續資料，最常使用迴歸分析。

故在迴歸分析時，可同時跑點陣圖/散布圖（Scatter grams），觀察是否有偏頗資料，如果有，應予刪除調整。

五、非線性和異質性處理

非線性（Nonlinearity）和異質性（Heteroscedasticity）,前者係就一個向面觀測兩組以上變項,彼此不呈線性;後者係同時兩個向面觀察兩組以上變項,彼此不呈線性。研究變項如果出現這兩個特性且特性十分強烈,將有損基於線性假設的分析結論。

這兩種特性通常也均可由閱讀雙變項的點陣圖/散布圖（Scatter grams）判別出來。

六、相依性和一向性

相依性或譯共線性（Multicollinearity）和一向性（Singularity）其實是一體的兩面,意指兩組以上的變項,是否僅為名稱不同,而在測量同一特質。相依性是指兩個以上變項形成同一模式,一向性則反過來說,一個單獨變項可以預測一組模式中其他的變項。如果發生嚴重相依性和一向性,則應剔除若干變項,不宜視為不同變項處理。

相依性和一向性可由相關矩陣中觀測而出,如果矩陣中有一對以上的自變項相關達到0.8至1.0 之間,這兩個變項便不宜作為不同自變項處理（Jae-On Kim & F.J.Kohout,1978）。

在多變項分析系列中，將就各別統計工具，再作詳細申述。