試題反應理論的介紹(五)

——模式與資料間適合度的檢定

英譯：(The assessment of model-data fit)

余民寧　著

英譯名：(Yu Min-Ning)

試題反應理論的特性與優點已在前幾篇文章中介紹過了，這些特性與優點並不是隨時都存在，它們只有在所選用的某種試題反應模式能夠適用某種感興趣的測驗資料時，才能夠存在；換句話說，在使用試題反應理論時，我們必須先檢定模式與資料間是否具有滿意的適合度(goodness-of-fit)，以確定所選用的模式能夠適用於所分析的資料，方不致於誤用或濫用試題反應理論的特性與優點。

檢定模式是否能適用於所分析資料的方法有許多種，Hambleton & Swaminathan(1985)建議從下列三方面來作爲判斷的依據：

1.模式對資料所具有的基本假設是否能夠滿足？

2.模式所具有的特性（如：試題與能力參數的不變性）是否能如期獲得？

3.在使用真實和模擬資料下，模式預測力的正確性爲何？

從上述三方面來進行模式與資料間的適合度檢定，可以幫助試題反應理論的使用者慎選適當的模式，作爲應用試題反應理論的先前準備。以下便從上述三方面來介紹常用的檢定方法。

<u>模式假設的檢定</u>

我們可以根據不同的模式所具有的不同假設來進行檢定。比較常見的檢定假設和方法計有：

一、單向度假設的檢定

1.根據試題與試題間相關係數矩陣來進行因素分析，再依特徵值(eigenvalues)大小，依序畫出特徵值分佈圖，再判斷該圖是否有一個明顯的主要因素存在。

2.比較真實測驗資料與隨機測驗資料（樣本數與試題數均相同）二者的試題間相關係數矩陣所畫成的特徵值分佈圖，如果單向度假設成立的話，則除了真實資料中的第一個特徵值外，這兩個特徵值分佈圖應該會很相似，而真實資料的第一個特徵值應該會比隨機資料的第一個特徵值還來得大。

3.檢查考生在能力量尺或測驗分數量尺的不同範圍內，其變異數——共變數矩陣或相關係數矩陣的局部獨立性假設。當單向度假設（大約）成立時，該矩陣的非對角線元素值會很小，且趨近於零。

4.針對試題間相關係數矩陣進行非直線的一個因素分析模式的因素分析，以檢定它的殘差值，並判斷是否仍有其他因素存在。

5.利用一種直接以試題反應理論爲基礎的因素分析的方法，來檢定測驗資料是否具有單向度的可能(Bock, Gibbons & Muraki, 1988)。

6.檢查某些看起來像是會違反假設的試題，看看它們是否表現出不同的功能。我們可以分測驗的形式和總測驗的形式，分別計算出這些試題的*b*值，如果單向度假設

成立的話，這兩種形式所計算出的*b*值所畫成的圖，應該會呈現直線分佈的情形，並且具有可資比較的試題參數估計值的標準誤。

除了上述的檢定單向度假設所採用的方法外，Hattie(1985)曾提出八十八種指標，作爲檢定單向度假設的參考，他結論認爲這些古老的心理計量學文獻所提供的檢定方法，多數都無法獲得令人滿意的結果，唯有以非直線的因素分析和殘差值分析爲基礎的方法，才能獲致最令人滿意的檢定結果。上述六種方法即是最具潛力的幾種，其他可能的方法也正在發展中。

二、相等鑑別度指標假設的檢定

這個假設檢定通常僅適用於一個參數模式，因爲它假設每個試題的鑑別度指標都相等。

我們可以從一種標準的試題分析中，逐題檢視試題與測驗分數間相關係數（二系列相關或點二系列相關係數）的分配，如果每個分配都呈現同質形狀時，我們所選用的模式便算符合相等的試題鑑別度假設。

三、最小猜測度假設的檢定

這個假設檢定通常也只適用於一個和二個參數模式，因爲它們均假設猜測度的可能性是微乎其微，甚至於完全沒有。檢定的方法至少有下列三種：

1.我們可以檢查低能力組考生在最困難試題上的表現情形，如果他們的表現水準是趨近於零，則這個假設可算是獲得滿足。

2.我們也可以訴諸試題與測驗分數間的迴歸線圖的幫助。測驗得分低的考生若傾向有接近於零的表現水準，則這個假設亦算是獲得滿足。

3.我們也可以檢視測驗難度、時間限制、與試題的編排格式等，以檢定猜測對測驗表現的可能影響力。

四、非速度（難度）測驗假設的檢定

這個假設和單向度假設一樣，均適用於所有的試題反應模式。我們可用至少下列三種方法之一來加以檢定：

1.我們可以比較沒有回答的試題數之變異數和答錯的試題數之變異數，當這個假設滿足時，這項比值應該是接近於零。

2.我們也可以比較在有時間限制下和沒有時間限制下的考生測驗分數，如果這兩次考試的表現情形具有高度的重疊部份，則表示這個假設獲得滿足。

3.我們也可以比較答完全部試題的考生百分比、答完百分之七十五試題的考生百分比、和被百分之八十考生答完的試題數，當幾乎所有的考生答完幾乎所有的試題時，速度便可被判定爲不是影響測驗表現的一個重要因素。

<u>模式特性的檢定</u>

最常檢定的兩種模式參數的特性爲：能力參數的不變性和試題參數的不變性。

一、能力參數估計值的不變性之檢定

我們可以拿不同測驗試題樣本所得的能力估計值來作比較（例如：比較困難與簡單的試題，或由題庫中抽取不同內容範圍所組成的測驗所估計出的能力參數估計值）。如果該估計值所相對應的測量誤差間差異不大時，不變性的特性便算是符合。

二、試題參數估計值的不變性之檢定

我們可以比較兩組或多組受試者（例如：男人和女人；黑人、白人、和西班牙裔人；教學組別；高分與低分的考生；不同地區來應考的考生等）接受某種測驗後，所獲得該測驗的試題參數估計值（例如：b 值、a 值、或 c 值）。根據兩組參數所畫成的分佈圖，除了因樣本大小所造成的分散誤差外，這圖應該是呈直線分佈，且基準線是由兩個隨機的相等樣本所建立，若此，則參數估計值的不變性才算存在。

總之，兩組模式參數（即根據同一批受試者在兩種測驗試題上的反應資料，所求得之能力參數估計值，和同一批測驗試題讓兩組受試者施測後，所求得之試題參數估計值）所畫成的分佈圖，可用來判斷該分佈圖是否呈直線分佈情形，若呈近似斜率爲 1，截距爲 0 的直線，則可說是某個試題反應模式適用於該份測驗資料，且具有模式參數不變性等特性。

<u>模式預測力的檢定</u>

另一種檢定模式與資料間適合度的作法，便是進行試題殘差值的分析。我們可以挑選一個合用的試題反應模式，並且估計出試題與能力參數，和求出各種不同能力組考生的表現情形，接著就可以比較預測的結果和真實的結果(Kingston & Dorans, 1985)。

某組考生在實得的試題表現(observed item performance)與期望的試題表現(expected item performance)之間的差距，便叫作原始殘差值(raw residual)，記作 r_{ij}。其數學公式如下：

$r_{ij} = P_{ij} - E(P_{ij})$（公式一）

其中，i 代表試題，j 代表某組考生的能力組別，P_{ij} 便是第 j 個能力組別在第 i 個試題上正確反應的實得百分比，而 $E(P_{ij})$ 則是在所選定（假設）的試題反應模式下正確反應的期望百分比。我們可以估計出假設的模式參數估計值，再利用這些估計值去計算一個正確反應的機率，這個機率便用來作爲某個能力組別的正確反應的期望百分比。

使用原始殘差值有個缺失，那就是無法顧及某個能力組別內期望的百分比正確分數的抽樣誤差。爲了顧及這項誤差，我們可以將原始殘差值除以期望的百分比正確分數的標準誤，以將原始殘差值轉換成標準化殘差值(standardized residual) Z_{ij} 如下：

$$Z_{ij} = \frac{P_{ij} - E(P_{ij})}{\sqrt{E(P_{ij})\,[1 - E(P_{ij})]\,/\,N_j}}$$ （公式二）

其中 N_j 是在能力組別爲 j 的考生人數。

當我們選擇試題反應模式時，原始殘差值、標準化殘差值、或兩者的分析，可以提供許多參考的訊息。下列便是檢定模式預測力常用的方法：

1.檢查模式與資料間適合度之殘差值和標準化殘差值。決定模式是否具有適合度，有助於挑選一個令人滿意的試題反應模式(Ludlow, 1985, 1986)。

2.在假設所有的模式參數估計值都正確的前提下，我們可以比較實得的與期望的測驗分數的分配，卡方統計數（或其他統計數）或圖解法可用來呈現這種比較結果。

3.我們也可以檢驗試題替換的影響、練習的影響、測驗的速限和作弊的影響、疲勞、課程、模式選擇不當、指導語的前後時效、認知處理的變項、以及其他會違害試題反應理論結果效度的不良影響，並且利用這些證據作爲挑選某種適當的試題反應模式的參考。

4.畫出能力估計值和其相對應的測驗分數間的資料分佈圖。當適合度的指標落在可被接受的範圍內時，除了少數的資料點在測驗特徵曲線（反映出測量誤差）周圍作零星分散外，該資料分佈圖應該呈現出強烈的直線關係才對。

5.運用許多統計考驗的方法來檢定整個模式、試題、或個別受試者的適合度。

6.使用電腦模擬的方法來比較真實的與估計的試題與能力參數。

7.利用電腦模擬的方法來檢定模式的韌性(robustness)，例如，我們可以研究單向度的試題反應模式能否適用於多向度的資料(Ansley & Forsyth, 1985; Drasgow & Parsons, 1983)。

實際說來，爲了計算殘差值，我們通常把能力量尺分割成等距的段落 10 至 15 個區間。這些區間必須要夠寬，以免落在這區間內的考生數過少，因爲樣本數過少所得的統計數會不穩定；同時，這些區間也必須夠窄，才能使得落在這區間的考生在能力上是屬於同性質的。

接下來是計算實得的百分比正確分數：算一算在某一能力組別內的考生答對某試題的總數，再除以該能力組別內的考生總人數。同時，習慣上是以每一能力組別的組中點來代表該組別的能力參數值，然後以該值來計算某個正確反應的機率，並求出每一能力組別中每一位考生在某個正確反應上的機率，這些機率的平均值即當作該能力組別的期望百分比。有了實得的百分比正確分數和期望的百分比正確分數之後，我們便可代入公式二進行某種統計考驗。

常用的統計考驗方法是卡方考驗(chi-square test)。Yen(1981)提出的 Q_1 指標，便是一種典型的卡方考驗所用的統計數指標，它可以用來檢定模式是否適合資料。某個試題 i 的 Q_1 統計數爲：

$$Q_{1i} = \sum_{j=1}^{m} \frac{N_j\,[P_{ij} - E(P_{ij})]^2}{E(P_{ij})\,[1 - E(P_{ij})]}$$ （公式三）

$$= \sum_{j=1}^{m} Z_{ij}^2$$

其中，根據能力估計值的不同，考生共可分成 m 個能力組別。Q_1 統計數將成爲一種以 $m-k$ 爲自由度的卡方分配，其中的 k 即是試題反應模式的參數個數。如果所計算出的 Q_1 值大於所查表的臨界值，我們便可以推翻試題特徵曲線（或試題反應模式）適合資料的虛無假設，而該建議尋找另一個較佳的模式才對。

總之，一個檢定模式與資料間適合度的方法，最理想的是包含(a)設計和執行各種分析，以便檢查不適合情況的可能型態，(b)仔細考慮通盤的結果，(c)根據所欲應用的範圍，判斷模式是否合適。而分析的過程應包括對模式的假設、模式的特性、和模式預測力與實際資料間的差異等之檢定，之後，再用統計考驗的方法來檢定虛無假設是否成立，以便提供統計訊息，作爲挑選一個適當模式的參考。

參考書目

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.

Hattie, J. A. (1985). Methodological review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. Applied Psychological Measurement, 9, 281-288.

Ludlow, L. H. (1985). A strategy for the graphical representation of Rasch model residuals. Educational and Psychological Measurement, 45, 851-859.

Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. Applied Psychological Measurement, 10, 217-229.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.