傾向分數（Propensity Score）— 非隨機對照研究之偏差消減（Bias Reduction）法

（中國醫藥大學附設醫院心臟內科）何弘棋 醫師

介紹

在觀察性研究中，研究人員無法控制治療之指派 (Treatment Assignment)。治療組和對照組在可觀測的共變數 (Covariates) 中可能具有很大的差異，這些差異可能導致療效差異估計之偏差。即使是傳統的共變異量分析 (Covariance Analysis) 調整也不一定能消除這種偏差。傾向分數，定義為受試者被指派到治療組之機率當給定共變數時，可用於平衡兩組間具有差異的共變數，從而減少估計之偏差。為了估計傾向分數，我們必須對治療指派建模當給定共變數時。一旦被估計，傾向分數可被用於匹配 (Matching)，分層 (Stratification)，迴歸調整 (Regression Adjustment)，或前三者的某種組合 , 以達到偏差消減的目的。本文即根據上述用途加以說明。[1]

傾向分數法已被廣泛用於許多科學領域，例如醫學，流行病學，衛生服務研究，經濟學和社會科學等。在隨機試驗中，隨機的治療指派保證了不同組別之間不存在共變數的系統性差異。 然而，在非隨機的觀察性研究中，研究人員無法控制治療指派，因此直接比較不同組別的結果可能會產生誤導。如果將共變數的信息納入研究設計 (例如，通過匹配抽樣) 或治療效果的估計（例如通過分層或迴歸調整），則可以部分避免這種缺陷。傳統的調整方法（匹配，分層和迴歸調整）往往受到限制，因為它們只能針對有限數量的共變數進行調整。然而，傾向分數不受此限地將共變數的信息統整為一個純量 (Scalar) 摘要。形式上，傾向分數定義為個體被指派接受治療的機率當給定共變數時。直覺上，傾向分數等同於個體接受治療的概度 (Likelihood) 當給定共變數時。Rosenbaum and Rubin 的研究顯示傾向分數是一平衡分數，可經由上述的方法用於觀察性研究以減少偏差。

定義

對於完整資料，Rosenbaum and Rubin 將個體 $i(i = 1,\cdots,n)$ 之傾向分數，$e(x_i)$，定義為此個體被指派為治療組 ($Z_i = 1$ *vs.* $Z_i = 0$) 之條件機率當給定共變數 $X_i = x_i$ 時：

$$e(x_i) = \Pr(Z_i = 1|X_i = x_i).$$

基本假設是，當給定 X_i's 時 , Z_i's 彼此相互獨立：

$$\Pr(Z_1 = z_1,\cdots,Z_n = z_n|X_1 = x_1,\cdots,X_n = x_n) = \prod_{i=1}^{n} e(x_i)^{z_i}(1 - e(x_i))^{1-z_i}。$$

傾向分數是所有平衡分數中最粗糙的 (Coarsest), 意即它是所有平衡分數的函數。而

平衡分數則定義為任一共變數的函數 $b(X)$，使得在治療組 $(Z = 1)$ 或對照組 $(Z = 0)$ 中 X 的條件分佈是相同的當給定 $b(X)$ 時。對於一傾向分數的特定值，治療組中所有具該傾向分數的個體的療效的平均值與對照組中相對應的平均值的差異是在此傾向分數的療效差異的無偏估計值 (Unbiased Estimate), 當治療指派，Z，是強可忽略 (Strongly Ignorable) 時。此概念可量化表達為：

$$E[r_1|Z = 1, b(X) = b] - E[r_0|Z = 0, b(X) = b] = E[r_1|b(X) = b] - E[r_0|b(X) = b],$$

其中 r_1 是治療效果，r_0 安慰劑效果。強可忽略被定義為 Z 與反應變數，Y，彼此條件獨立當給定 X 時，也就是 $Y \perp Z|X$。因此，借助傾向分數的匹配，分層或迴歸調整往往可獲得療效差異的無偏估計當治療指派是強可忽略時。

當共變數不包含丟失 (Missing) 數據時，傾向分數可以藉由判別分析 (Discriminant Analysis) 或邏輯迴歸 (Logistic Regression) 來估計。這兩種技術都可估計治療指派的條件機率當給定共變數時。形式上，當使用判別分析時，共變數被假設為服從常態分佈，但邏輯迴歸則不需該假設。

未曾使用過傾向分數的研究人員可能會有這樣的疑問：“為什麼我們必須估計受試者接受治療的機率？既然我們已經確知受試者接受了哪種治療。”這個問題的答案是，如果我們使用受試者接受治療的機率 (即傾向分數) 來調整我們對療效差異的估計，我們可以創建一個“擬隨機化 (Quasi-randomized)”試驗。也就是說，如果我們發現治療組與對照組中各有一個受試者具有相同的傾向分數，那麼我們可以想像這兩個受試者是被“隨機指派”的，意味他們被指派到治療組或對照組的機率相同。在隨機對照試驗中，隨機分組將個體指派給治療組和對照組，優於藉由傾向分數所做的調整，因為它不依賴於特定的共變數，而是適用於任何已觀察到或未觀察到的共變數。雖然使用傾向分數所得的結果僅有條件地有效當給定已觀察到的共變數時，但是如果研究者有能力測量許多被認為與治療指派相關的共變數，則可以很有把握地得到一個無偏估計。

傾向分數的使用

目前在觀察性研究中，傾向分數主要用於減少偏差和提高精度 (Precision)。最常利用傾向分數的技術是匹配，分層和回歸調整。這些技術分別在計算療效差異之前，之前與當下，和當下對共變數進行處理。這三種技術求取傾向分數的計算方式相同，但應用方式不同。傾向分數有利於這些技術的運用，因為傾向分數定義為個體被指派到治療組的機率（亦即，$e(X) = \Pr(Z = 1|X) = E[Z|X]$）當給定共變數時，其暗示 Z 和 X 彼此條件獨立當給定 $e(X)$ 時。此性質的量化理由是：

$$e(X) = E[Z|X] \Rightarrow e(X) = E[Z|e(X)],$$

其表示當給定 $e(X)$ 時，個體被指派到治療組的機率就是 $e(X)$，一個常數，而常數跟任何隨機變數互相獨立。因此，具有相同傾向分數的受試者，在治療組和對照組之間，其共變數

往往具有相同的分佈。藉由傾向分數所做的精確調整將平均地去除背景共變數在組別之間的偏差，因此可以統括性地使用傾向分數 (一個純量)，而不是逐一使用所有背景共變數去做去除偏差的處理。

匹配

研究人員經常面臨的狀況是治療組的人數有限，但對照組的人數卻相當龐大。一個例子是由出生缺陷基金會 (March of Dimes) 所資助的研究，探討過期分娩 (Post-term Birth) 對學齡兒童 (5-10 歲) 在神經精神，社會和學術成就方面的影響。在研究開始時，研究人員收集了超過 9000 份出生記錄 (749 名治療組 (過期) 嬰兒和 9000 多名潛在的對照組 (足月) 嬰兒)，並提供產前和出生病史資料。因經費有限，研究人員不可能對所有對照組兒童進行採樣，因此必須進行某種形式的抽樣。

匹配是一種常用的技術，用於選取某些對照組受試者，以便控制治療組與對照組之間背景共變數上的差異。雖然找匹配的想法是直截了當的，但其困難點是，即使只有少數幾個背景共變數，也不易找到匹配的對象同時具有類似的共變數。出生缺陷基金會的研究人員必須面對這個問題，因為他們有超過十個共變數需要匹配。

傾向分數匹配法解決了這樣的問題，其藉由匹配一個純量，允許研究人員同步匹配許多背景共變數。傾向分數匹配法之前，一個常用的方法叫馬氏匹配法 (Mahalanobis Metric Matching)。馬氏匹配法隨機排序受試者，然後計算第一個治療組受試者與所有對照組受試者之間的距離，其中治療組受試者 i 和對照組受試者 j 之間的距離標示為 $d(i,j)$。具有最小距離的對照組受試者 j 即被選為治療組受試者 i 的匹配對象，此兩者即被移除，不參與下一次的匹配。此過程不斷重複，直到找到所有治療組受試者的匹配對象。這種技術的一個缺點是，當模型中包含許多共變數時，很難找到緊密的匹配對象。而隨著共變數個數增加，個體之間平均距離也會增加。另一方面，傾向分數可以使用許多共變數來計算，但其分數仍然是一個純量摘要，因此匹配通常很容易。

分層

分層 (有時也稱為亞分類)，也常用於觀察性研究，以控制治療組和對照組之間的系統性差異。該技術基於背景共變數，將受試者分層。當各分層被決定後，屬於同一分層但不同組別的受試者，即被直接比較。當共變數個數增加時，許多在匹配時會發生的問題，也會隨之發生。科克蘭 (Cochran) 指出隨著共變數個數的增加，層數呈指數增長。例如，如果所有共變數都是二元變數，則 k 個共變數將存在個分層。如果 2^k 太大，那麼一些分層可能僅包含治療組受試者，這將無法估計該分層的療效差異。這時傾向分數法就非常有用。因為傾向分數法可獲取一純量摘要，層數不會隨共變數個數的增加呈指數增長。

Rosenbaum and Rubin 的理論結果顯示 , 基於傾向分數的完美分層將保證層內的療效差異是真實療效差異的無偏估計。再一次，他們

假設治療指派是強可忽略的。Rosenbaum and Rubin 在基於傾向分數的分層法中，得到一致科克蘭的結果：創建五個分層可移除 90% 的偏差，他們還表示，事實上，傾向分數分層法不但可平衡用來估計傾向分數的 k 個共變數，而且五層傾向分數分層法經常可移除 90% 的偏差。

用於確定分層的技術是直接了當的。首先，藉由邏輯迴歸或判別分析估計傾向分數。研究者然後必須決定分層間的邊界值。我們通常用全體受試者的傾向分數的五分位數來當成不同分層的邊界值。

例如，研究員史東希望了解 747 例社區性肺炎 (CAP) 的患者的預後，是否跟住院與否有關 (n=265 vs. n=482)。 由於患者未被隨機指派為住院或非住院，因此使用分類樹技術來估計傾向分數。 接著患者依傾向分數被分成七層。研究人員發現，原本 44 個共變數中有 29 個存在兩組之間的不平衡，而在傾向分數分層之後，減為 13 個，然後研究人員使用直接標準化方法估計分層特定 (Stratum-specific) 的療效差異。

迴歸 (共變異量) 調整

傾向分數也可用於迴歸調整。在迴歸調整中，療效差異，τ，可以被估計為

$$\hat{\tau} = (\bar{Y}_t - \bar{Y}_c) - \hat{\beta}(\bar{X}_t - \bar{X}_c)，$$

其中 Y 是療效，t 治療組，c 對照組，$\hat{\beta}$ 共變數 (X) 迴歸係數的估計。上面的等式通過減去右側第二項的共變數效果來執行迴歸調整。由於上述的理由，傾向分數是迴歸調整中的有用變量，因為研究者只要將傾向分數放入迴歸分析中，就可求得其迴歸係數的估計，而達到迴歸調整的目的。Rosenbaum and Rubin 發現只要治療組和對照組的反應面 (Response Surfaces) 是平行的，不論線性或非線性，則使用傾向分數的迴歸調整最終都可降低療效差異估計的偏差。此外，如果在一個特定分層內使用傾向分數迴歸調整，那麼療效差異的估計將比單用匹配來得有效率。迴歸調整的另一種作法是先使用大量背景共變數來估計傾向分數，然後將這些共變數的的一個子集和傾向分數放入最終的迴歸分析中。

研究員 Muller 等人希望了解毛地黃對心肌梗塞後患者死亡率的影響。這裡非隨機對照治療是毛地黃。研究人員根據 19 個共變數計算出一個“不平衡風險分數”，它似乎是傾向分數。該模型中的共變數包括受試者的心率，年齡，以及受試者在前三周是否有用 β- 受體阻斷劑。考慮了基線預後因子的影響，他們使用 Cox 比例風險迴歸來決定毛地黃與存活之間的相關性。似乎他們通過在其模型中導入傾向分數來考慮基線差異，以便調整其最終療效差異的估計。

使用傾向分數執行迴歸調整時可能出現的一個問題是，使用傾向分數是否真有增益，相對於使用所有的共變數而言。執行兩步程序的一個優點是，可以使用含較多交互作用和較高階項的複雜模型來估計傾向分數。由於傾向分數的目標是獲得治療指派機率的最佳估計，所以不擔心過度參數化的問題。然後，當傾向分

數得到估計後，研究人員可以在迴歸分析中僅包含少數最重要的共變數和傾向分數。這種較小的模型允許研究員更可靠地做模型檢查。

一般來說，應該謹慎地執行迴歸調整。Rubin 指出如果治療組和對照組共變異量矩陣不相等，則迴歸調整實際上可能會增加偏差平方的期望值。另一個困難可能會發生在當治療組和對照組的方差非常不同時 (即，對照組的方差遠大於治療組的方差)。在這些情況下，可以考慮使用傾向分數進行匹配或分層，而不是進行迴歸調整。

討論

在上述一些例子中，並沒有提到當資料有缺失時，該如何估計傾向分數。這是大多數實務應用中的一個重要問題。例如，出生缺陷基金會贊助的資料中也有缺失的問題。目前，正有許多方法被陸續開發以處理不同的缺失機制。另一個正被研究的領域是使用傾向分數來估計臨床試驗中的療效差異，當受試者可能提早退出試驗時。在這裡，傾向分數被定義為完成試驗的機率當給定基線和早期結果時。

傾向分數在統計分析中被廣泛使用，特別是在應用醫學領域。隨著隨機對照試驗的成本不斷上升，更多的研究人員轉而使用觀察性研究作為進行較便宜研究的手段。傾向分數最大的增益是在當可以被融入研究設計階段 (通過匹配或分層) 時。這些好處包括對真正的療效差異做更準確的估計，並節省時間和金錢。這種節省是出於能夠避免招募可能不適合特定研究的受試者。最後，本文並不主張在觀察性研究的分析中僅使用傾向分數，而是鼓勵使用傾向分數以擴充傳統的分析方法。在研究中，傾向分數應該被視為研究人員可以使用的額外工具，因為他們試圖估計研究中的療效差異。

參考文獻

1. d'Agostino, R.B., *Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group.* Stat Med, 1998. **17**(19): p. 2265-2281.