傾向分數(propensity score)在估計風險比之使用方法

陳錦華 資深統計分析師/副教授

計算傾向分數的目的

在收集資料類型中，常用的一種是觀察型研究(observational studies)，用此來估計治療、介入及暴露對結果(outcome)之影響。雖然隨機對照試驗可以控制外來的干擾因素，但非所有的研究皆適用，如：想了解吸菸對於疾病的影響，則無法強迫受試者暴露於吸菸的狀態中，直到研究結束，如此作法不合乎倫理規範。故需使用觀察型研究，使達到探討此議題的目的，此研究設計的缺點，是無法讓吸菸及不吸菸的兩群人在研究開始之條件(baseline，或稱基準)都控制得差不多，如：兩組年齡、性別、共病症之狀態…等等(這些變項為潛在的干擾因子(confounder))。若相同的話，進行比較時就能排除這些因素的影響，製造”很像”隨機對照試驗的研究設計，在比較治療效用或暴露效用時，則能準確估計兩組之差異。為了要消除這些干擾因子的影響，或使其影響減少到最小，進而達到準確估計的目的，可利用傾向分數來達成。常使用的方法如下(以下討論以治療效用為主要探討的變項)：

1. 根據傾向分數進行配對
2. 利用傾向分數進行分層
3. 利用傾向分數倒數對治療組及非治療組進行加權
4. 將傾向分數當成一個變項，在模型中控制(adjusted)

當反應變項(response variable or outcome)為連續時，以上四種處理方式對於治療效果的估計皆是不偏估計[(1)]，也就是可達到”類似”隨機對照試驗的設計，將干擾因子作適當控制，以估計正確的治療效果。不過，在醫學的研究上，反應變項通常是二元變項(binary, 如：得病或沒病)、或觀察到事件發生時間(time-to-event)。本篇文章中，我們將針對”觀察到事件發生時間”為反應變項，上述四種方法是否能適當的減少風險比(HR)的估計偏差。

傾向分數的計算

傾向分數：為特定個體會接受治療的機率，範圍在 0-1 之間。治療與否為二元變項(令治療與否為 Z=1 或 Z=0)，可利用邏輯斯迴歸模型(logistic regression model)估計傾向分數。但需要決定選擇治療與否的變數(**X**)為何，即治療與否受到哪些變數影響，通常會考慮常見的干擾因子（如：年齡、性別、共病症....等等），可將這些變項視為解釋變項(explanatory variable)放入邏輯斯迴歸模型之內，將多個變項轉換為一個單一個分數，來表達接受治療之機率。

傾向分數：$e(\mathbf{X})=\Pr(Z=1|\mathbf{X})$，

$$\log\left(\frac{e(X)}{1-e(X)}\right)=\alpha+\beta_1 x_1+\beta_2 x_2....+\beta_k x_k \ ,$$

傾向分數主要目的是讓治療組和非治療組在干擾變項(**X**)的分布大致相同，平均上讓研究開始時之兩組能達到類似的基準條件，像是隨機對照試驗那樣，使後續對於兩組之結果評估具公平性，在控制干擾因子下，正確估計出治療效用，產生無偏差之估計。Rosenbaum[(2)]建議，在研究開始時(baseline)，評估治療組及非治療組在研究開始前一年或半年，其干擾因子之狀態，在哪些變項(干擾因子)具有差異，考慮將這些具差異的變項納入傾向分數之計算。不過，這分數的計算仍存有主觀因素存在。Kurth et al.[(3)]研究：在住院期間使用 t-PA 治療缺血性中風病人之死亡率評估，主要想看治療與否對死亡之影響。此研究中，在影響治療與否變項為：年齡、性別、從產生中風徵兆至住院之時間、住院次數、局部麻痺的程度...共 16 個變項，並計算傾向分數。不過，此研究開始時，有些變項在治療組及非治療組仍是無統計上顯著差異，仍被納入傾向分數中。

傾向分數應用於風險比(hazard ratio)的估計

若反應變項為”觀察到事件發生時間(time-to-event)”，此事件可能是死亡、復發、癌症發生...等情形。通常用存活分析方法，即估計 Cox model 中之係數，轉換後為風險比(hazard ratio)，來解釋治療效用對於死亡(或復發、癌症發生...)的影響。此

時，傾向分數調整方法不同，所得係數估計之準確度會受影響。常見的問題：使用不適當的統計方法及無法正確評估傾向分數包含之變項是否能平衡治療組及非治療組在研究開始之差異。

如同隨機對照試驗一樣，傾向分數是用於估計邊際效用(marginal effect，指對於整個母體之平均效用，即模型中只有治療與否之變項)，並不用於估計條件效用(conditional effect，對於特定個人之效用，即考慮多變項模型，治療與否只是其中一個變項)[4]。這裡主要針對邊際風險比(marginal HR)之討論，如隨機對照試驗中，針對治療與否對死亡或疾病之影響，只著重邊際風險比估計。下以說明使用之調整方法。

1. 根據傾向分數進行配對

治療組和非治療組以相同或接近的傾向分數予以配對(1:1 或 1:N)。在配對後，每成對非為獨立，大多忽略此關係而使用傳統 Cox model。另一方面，若要檢視治療組及非治療組 Kaplan-Meier survival curve 的差異時，需視兩組為獨立，利用 log-rank test 檢定，也是不當的作法，可用 stratify log-rank 檢定予以修正。利用 Cox model 估計的方法有下列三種的比較：

I. naive Cox model：以傳統 Cox model 估計治療效用，並利用模型結果估計 95% 信賴區間。雖然忽略成對資料具相關性考量，不過在估計風險比上仍是不偏估計值，但會錯估係數的標準誤，影響檢定結果。

II. robust Cox model：以上述方法估計治療效用，以穩健三明治估計法(robust sandwich estimate)得到迴歸係數之標準誤，可衡量資料群聚相關性，以得到 95%信賴區間。此法所估計邊際風險比為不偏(unbiased)估計值，估計值之變異較小。

III. stratified Cox model：以配對後之樣本，根據傾向分數分層，以單變項 Cox model 估計每分層下之風險比，再針對這些值予以比較。此法所得之邊際風險比為偏差(biased)估計值。

2. 利用傾向分數進行分層

用百分位數將傾向分數分成五等份，每等份具相同資料個數，此為分層之依據。在估計線性治療效用(linear treatment effect，如：平均數差或比例差，反應變項是連續資料)時，可以消除接近 90%之偏差，可算是不錯的調整方法；利用各分層下所估計出來之線性治療效用後，再加以平均或合併以得到最終之估計治療效用。

但若估計對象是風險比而非線性治療效用呢？通常考慮以下三種模型：

I. stratification adjusted：利用 Cox model，在解釋變項中，定義治療與否為二元變項，傾向分數以五個類別變項直接在模型中調整(只有四個虛擬變項(dummy variable))。

II. stratification-pooled：在五個不同分層下，估計單變項 Cox model 之治療效用(ln(HR))，並將五個分層下，所得到的治療效用予以合併或計算平均值，以估計治療之總效用。

III. stratification-stratified：如上述 II.之作法，但不計算合併後的總效用，在不同分層下，解釋其治療效用隨著分層不同之變化。

3. 治療變項利用傾向分數倒數予以加權

定義 IPTWs：權重 $w = Z/e + \left[(1-Z)/(1-e)\right]$，Z=1 為治療組，Z=0 為非治療組。將治療與否以傾向分數作加權，治療組的低傾向分數及非治療組之高傾向分數給予較大權重，反之則反，以估計平均治療效用(ATE)。若權重為 $w = Z + \left[e(1-Z)/(1-e)\right]$，治療組之權重皆為 1，非治療組之權重為 $e/(1-e)$，則為估計治療組之平均治療效用(ATT)。在加權樣本下，無論在 ATE 或 ATT 之估計，皆利用 Cox model，以治療效用為目標，並使用穩健之標準誤估計。此類估計方法皆可得到不偏之估計值。在 Kaplan-Meier survival curve 對於兩組比較時，可用 adjusted log-rank 檢定。

4. 將傾向分數當成一個變項，在模型中當調整(adjustment)

Cox model 中將包含兩個變項，一個為治療與否之二元變項，另一個為傾向分數。這種方式所得之治療效用為有偏差之估計值，不建議使用。

結論

Austin[5]利用 Monte Carlo 方法模擬傾向分數及存活資料，並設定不同強弱風險比(參數)，每筆資料為 10,000 個人，並重覆模擬 10,000 次，主要針對以上四種傾向分數調整方式，給予使用者建議：

A. 上述方法中，以配對(naave Cox model 及 robust Cox model)及加權方式(IPTWs)，可得到準確之邊際風險比估計值(即偏差最小)，當治療組之人數愈多，會更精準。其中，配對的兩方法中(1-I.及 1-II.)，有相同點估計值。

B. 依配對分層(1-III.)、依傾向分數分層(2-I.~2-III.)及將傾向分數當作調整變項(4)，皆得到偏差之邊際風險比估計值(bias 較大)。這些方法皆不建議用於治療組及非治療組在研究開始之差異的調整。

C. 雖然配對(naive Cox model 及 robust Cox model)及加權方式(IPTWs)，共四方法無偏差之估計發生，但在重覆資料模擬下，發現 naive Cox model 之估計值變異程度較大，檢定時也較保守；robust Cox model 估計值變異較小，在檢定上具有優勢。IPTW 加權方法，ATE 和 robust Cox model 相同，具有變異較小的優勢；而 ATT 估計值變異程度卻大於 naive Cox model。

提供以上傾向分數的調整方法，用於存活資料分析時，使治療組及非治療組有相當的基準，以利往後估計正確之風險比。

參考資料

1. Rosenbaum PR, Rubin DB. The central role of propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55.
2. Rosenbaum PR. Design of observational studies. Springer-Verlag: New York, NY, 2010.
3. Kurth T, Walker AM, Glynn RJ, et al. Results of Multivariable Logistic Regression,

Propensity Matching, Propensity Adjustment, and Propensity-based Weighting under Conditions of Nonuniform Effect. American Journal of Epidemiology 2005; 163: 262–270.

4. Rosenbaum PR. Propensity score. In encyclopedia of biostatistics, Armitage P, Colton T (eds). John Wiley & Sons: Boston, 2005; 4267-4272.
5. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Statistics in Medicine 2013; 32: 2837-2849.
6. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Statistics in Medicine 2014; 33: 1242-1258.