[研究新領域報導]

因果中介模型

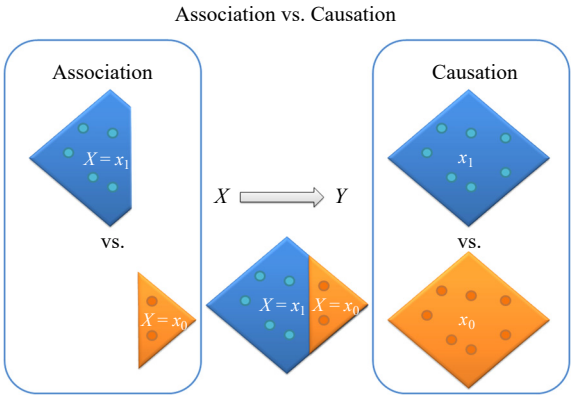
中央研究院 統計科學研究所　黃彥棕

一、因果推論

因果推論是一門以數理方法研究因果關係的學問，因果關係的確立仰賴所謂潛在結果（potential outcome 或 counterfactual outcome）的觀念，潛在結果的想法在統計文獻中，最早由波蘭數學及統計學家 Jerzy Neyman 在 1923 年的一篇關於隨機農業試驗分析的文章中所提出。美國統計學家 Donald Rubin 在 1974 年更將潛在結果的想法推廣至觀察型研究，此一架構更進一步由美國統計學家 James Robins 在 1986 年延伸至多重隨時間變化的因子，另外，潛在結果架構其背後的理論基礎亦被發現，與資訊科學以圖像理論爲主的因果推論(Spirites, Glymour and Scheines 1993; Pearl 1995)有共通之處。

假設我們想研究X是否造成Y的發生，一般的資料所觀察到的是某一個人「X出現，Y會發生」，而因果關係的確定需要的論證是，這個人「X出現，Y會發生；X若不出現，Y也不會發生」，然而，這樣的資料常常無法取得。例如，X是抽煙，Y是死亡，若某人抽煙者死亡了，除非有時光旅行或人死得以復生的技術，否則我們無法得知他在不抽煙的狀況下，是否能活得較久。不過，由於統計仰賴的不是單一個體的推論，而是群體的推論，所以需要的論證其實稍微弱一些，亦即我們需要的資料是，這“群”人「X出現時($X = x_1$)，Y“比較常”發生；X若不出現($X = x_0$)，Y“較不常”發生」。如圖一所示，這時我們需要的試驗可以想像爲，讓這群人在研究開始的時間原點全部都接受$X = x_1$的治療或暴露，然後觀察Y的發生率，我們令這個發生率爲，$\mathrm{E}[Y(x_1)]$，然後再讓這群人全部回到時間原點，但這次接受$X = x_0$，然後再次觀察Y的發生率，$\mathrm{E}[Y(x_0)]$，而$\mathrm{E}[Y(x_1)] - \mathrm{E}[Y(x_0)]$就是這個問題的因果效應。上述的$Y(x)$就是潛在結果符號，它所代表的是在接受$X = x$的治療或介入時所能觀察到的$Y$。因此如同$Y$，$Y(x)$也是個隨機變數，但不同於$Y$的是，$Y(x)$不見得能從資料中直接取得。

一般資料所收集到的如圖一中所示，有部份的人接受$X = x_0$，部份的人接受$X = x_1$，資料分析常做的是$\mathrm{E}[Y|x_1] - \mathrm{E}[Y|x_0]$以代表$X$和$Y$的相關性。我們接著想問的是，相關性等同於因果關係嗎？或者$\mathrm{E}[Y(x_1)] - \mathrm{E}[Y(x_0)] = \mathrm{E}[Y|x_1] - \mathrm{E}[Y|x_0]$成立嗎？又或者，$\mathrm{E}[Y(x)] = \mathrm{E}[Y|x]$（$x = x_0$或$x_1$）成立嗎？其中$\mathrm{E}[Y|x_1]$代表的是圖一左鑽石形狀的藍色族群中$Y$的發生率，而$\mathrm{E}[Y(x_1)]$代表的是圖一右圖完整的藍色菱形族群中$Y$的發生率，因此$\mathrm{E}[Y(x_1)] = \mathrm{E}[Y|x_1]$需要的假設是，上述那兩個族群是相同可互換的(exchangeable)，亦即，我們可以用鑽石形狀的部份族群去填補缺掉的三角形以得到完整的菱形；同樣的，如果圖一左橘色三角形族群和圖一右完整的橘色菱形族群是可互換的，那麼我們也能得到$\mathrm{E}[Y(x_0)] = \mathrm{E}[Y|x_0]$。而可互換的假設以數學符號代表：$Y(x) \perp X$，即潛在結果不受現實中所接受的治療$X$所影響。這樣的假設可能被一些狀況所違背，例如，如果接受x_0多爲女性，而接受x_1多爲男性，則$Y(x) \perp X$不成立，我們無法做上述的族群互換或填補，那麼，$\mathrm{E}[Y|x_1] - \mathrm{E}[Y|x_0]$所估計出來的相關性就不等同於因果效應，這一種造成無法互換的機制我們稱爲干擾機制(confounding)。如果我們已經知道干擾因子“Z =性別”的存在，並假設$Y(x) \perp X|Z$成立：即潛在結果雖然受現實中所接受的治療X所影響，但如果分別在男性及女性中，潛在結果即不再受現實中所接受的治療X所影響。那麼，我們仍可估計因果效應$\mathrm{E}[Y(x_1)|Z] - \mathrm{E}[Y(x_0)|Z] = \mathrm{E}[Y|x_1, Z] - \mathrm{E}[Y|x_0, Z]$，而此一估計與上述的不同之處在於，它是限定在特定的干擾因子Z之下。



圖一　因果性(causation)和相關性(association)的區別。圖形中圓形小點代表 $Y = 1$ 的發生。

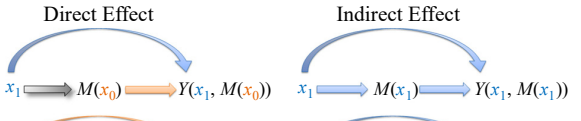
二、因果中介模型

除了X是否造成Y因果關係，中介分析想進一步探討X透過何種機制造成Y。中介分析最早由 Baron and Kenny 於 1986 年在社會科學文獻中所提出，中介分析將X對Y造成的作用分解爲：透過中介因子M所造成的間接效應(indirect effect)、以及不透過M造成的直接效應(direct effect)。在 1992 年時，美國統計及流行病學家 James Robins 及 Sander Greeland，以及 2001 年，資訊科學及統計學家 Judea Pearl 分別以潛在結果的架構下，研究了中介分析的直接及間接效應，並進行了嚴謹的因果推論，這些重要的研究爲統計學及因果推論確立了一門新興的領域：因果中介模型(causal mediation model)。首先，我們定義$Y(x_a, M(x_b))$爲接受$X = x_a$及$M = M(x_b)$雙重治療或介入的潛在結果，特別的是，$M = M(x_b)$的介入本身也是一個潛在結果，可以想像其爲另一組實驗所得到的一個機率分布，而我們依照這個機率分布給予介入的數值，利用這個雙重介入的潛在結果符號，我們定義直接及間接效應如下：

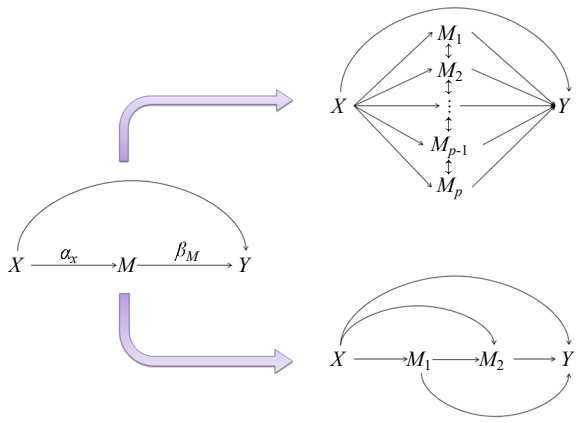
$$\Delta_{\mathrm{DE}} = \mathrm{E}\left[Y\left(x_1, M(x_0)\right)\right] - \mathrm{E}\left[Y\left(x_0, M(x_0)\right)\right]$$

$$\Delta_{\mathrm{IE}} = \mathrm{E}\left[Y\left(x_1, M(x_1)\right)\right] - \mathrm{E}\left[Y\left(x_1, M(x_0)\right)\right]。$$

而上述的定義亦可圖示爲圖二。近年更陸續有學者(VanderWeele and Vansteelandt 2009；Imai et al., 2010)提出因果中介分析所需要的辨識假設(identifability assumptions)及估計式。其所需的辨識假設簡言之即是上述可互換性的推廣，即X、M及Y之間沒有干擾機制的存在，另外亦需排除X可以透過其它因子同時造成M及Y。在最簡化的線性模型且無X與M的交互作用之下，間接效應的估計式可被表示爲一個簡單的乘積$\alpha_X\beta_M$，其中α_X代表的是X對M的因果效應，而β_M是M對Y的因果效應。此一簡單的結果可以幫助我們了解間接效應（或稱爲中介效應）的物理意義，它必需在X對M、以及M對Y同時存在有因果效應時，才會存在。若僅X影響M，但M不影響Y，這樣的現象並不會造成中介效應，這也相當符合直觀。



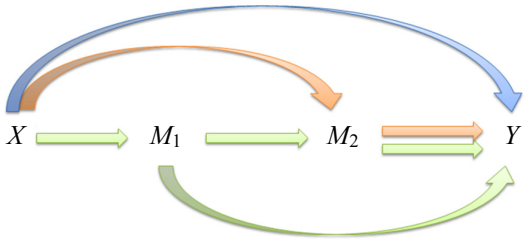
圖二　因果中介分析中的直接效應(direct effect)和間接效應(indirect effect)。



圖三　由單一中介因子的因果中介模型至多重中介因子的推廣。

1 多重中介因子模型

比起一般統計方法，因果中介模型能爲各種科學現象（例如疾病發生的致病機制、癌症的病程）提供更佳的詮釋。但是，傳統的中介模型著重在單一中介因子，但是許多科學研究中X和Y之間的關係，往往不能由單一中介因子所解釋。筆者的研究主要是將因果中介模型拓展至多重中介因子，如圖三所示，多重中介因子$\boldsymbol{M}^T =$



圖四　雙中介因子模型的特定路徑效應(path-specific effect)。可分為三種特定路徑效應：$X \to Y$，由X不透過中介因子到Y的效應（藍色路徑）；$X \to M_2 \to Y$，由X透過M_2到Y的效應（橘色路徑）；$X \to M_1 \to Y$或$X \to M_1 \to M_2 \to Y$，由X透過M_1到Y的效應（綠色路徑）。

$(M_1, M_2, \dots, M_p)$又分爲兩類，其一是縱向的多重中介因子，其二是橫向的多重中介因子。縱向多重中介分析著重的是X透過$\boldsymbol{M}$爲整體而造成Y而不去探究$(M_1, M_2, \dots, M_p)$之間的因果關係。筆者使用一個有效的降維方法，在縱向高維度多重中介模型中，即使在p（中介因子數目）大於n（樣本數）的情況下，亦可以有效檢定中介效應(Huang and Pan, 2016)。另一方面，橫向多重中介模型希望明確地將$(M_1, M_2, \dots, M_p)$之間因果關係納入分析，稱之爲特定路徑效應（path-specific effect，圖四），並希望知道X在這些中介因子的因果機制中，如何透過不同的特定路徑效應影響Y。筆者提出了一系列用於存活分析之多重中介模型，包括使用半母數概率模型(semiparametric probit model)、風險累加模型(Aalen's additive hazard model)及Cox風險等比模型(Cox proportional hazards model) (Huang and Cai, 2016; Huang and Yang 2017; Cho and Huang 2018+)。在這些存活模型中，我們推導出特定路徑效應的解析解。假設有兩個中介因子M_1及M_2，其挑戰在於特定路徑效應的推導仰賴以下積分：

$$\int_{m_1}\int_{m_2} F_{Y|M_2,M_1,X}(y|m_2,m_1,x_a)dF_{M_2|M_1,X}(m_2|m_1,x_b)dF_{M_1|X}(m_1|x_c)\ ,$$

其中$F_{Y|M_2,M_1,X}(y|m_2,m_1,x_a)$爲$Y$在給定$(M_2,M_1,X)$下的機率分布函數、$F_{M_2|M_1,X}(m_2|m_1,x_b)$爲$M_2$在給定$(M_1,X)$下的機率分布函數、$F_{M_1|X}(m_1|x_c)$爲$M_1$在給定$X$下的機率分布函數。上述的積分廣義而言可以使用重複選取的數值方法逼近，然而隨著中介因子數目的增加，數值逼近法需要大量的運算，並不符合實際應用。筆者在不同的模型下，推導出相應的解析解，或在無解析解的情況下，研究以其它模型逼近，並找出所需的假設條件。除了存活分析，我們也發展了當中介因子及結果皆爲二元的多重中介模型(Shih, Huang and Yang 2018)。

2. 因果中介假說檢定

因果中介分析的另一個挑戰是假說檢定，其困難之處在於它的虛無假說是複合性的：

$$H_0: \alpha_X\beta_M = 0$$

$$\leftrightarrow H_0^{(1)}: \{\alpha_X = 0 \cap \beta_M = 0\} \cup H_0^{(2)}: \{\alpha_X \neq 0 \cap \beta_M = 0\} \cup H_0^{(3)}: \{\alpha_X = 0 \cap \beta_M \neq 0\}$$

即上述所提到的，暴露(X)-中介因子（M）因果關係α_X、及中介因子(M)-結果（Y）因果關係β_M，兩者任一爲虛無效應$(=0)$，則中介效應爲虛無效應，這一類的虛無假設稱爲交集-聯集試驗(intersection-union test, Berger and Hsu 1996)。過去中介分析相關文獻中，除了以數值模擬的分析之外，尚未有研究嚴謹地處理此一問題。進行假說檢定有兩種主要的方法，假設α_X及β_M估計式分別爲$\hat{\alpha}_X$及$\hat{\beta}_M$，一種常用的檢定方法稱爲乘積檢定(product significance test)，其檢定統計量(T_{PT})爲著重於檢定乘積$\hat{\alpha}_X\hat{\beta}_M$；另一種檢定法爲聯合顯著檢定(joint significance test)，其檢定統計量(T_{JT})則分別檢定$\hat{\alpha}_X$及$\hat{\beta}_M$：

$$T_{PT} = \frac{\left(\hat{\alpha}_X\hat{\beta}_M\right)^2}{\hat{\alpha}_X^2 Var(\hat{\beta}_M) + \hat{\beta}_M^2 Var(\hat{\alpha}_X)}$$

$$T_{JT} = \min\left(T_\alpha^2 = \frac{\hat{\alpha}_X^2}{Var(\hat{\alpha}_X)}, T_\beta^2 = \frac{\hat{\beta}_M^2}{Var(\hat{\beta}_M)}\right)。$$

在虛無假說下，T_{PT}和T_{JT}皆服從自由度爲一的卡方分布，然而$T_{JT} \geq T_{PT}$的不等式結果造成以T_{JT}爲主的檢定將必然比T_{PT}擁有更高的統計效力(Huang 2018+a)。此外，筆者發現常用的T_{PT}服從卡方分布只有在α_X或β_M至少有一個參數不爲0的情況下才成立，亦即上述的$H_0^{(1)}$違背此一條件，而將造成T_{PT}檢定的誤差。筆者也證明了T_{JT}的檢定並不需要上述的條件，且其檢定規模(test size)爲α。由於T_{JT}的這些良好性質，筆者提出了以T_{JT}爲基礎的縱向及橫向多變量中介效應聯合顯著檢定，並證明這些檢定方法較乘積檢定有更高的統計效力，且檢定規模爲α (Huang 2018)。

由於中介效應的虛無假說包含三種不同類型的假設聯集，$H_0^{(1)} \cup H_0^{(2)} \cup H_0^{(3)}$，在三種類型的比例爲未知時，$p$值的運算有其困難。雖然可以證明假說檢定的檢定規模理論值爲α，在實際應用例如基因體的分析時，我們發現檢定規模理論值α很難達到，亦即，整個基因體的分析將發現$p < 0.05$的比例遠小於0.05。這是因爲檢定規模之定義爲$\sup_{\theta\in\Theta_0} \pi(\theta)$，其中$\theta = (\alpha_X, \beta_M)^T$、$\pi(\theta)$爲效力函數(power function)，而此函數是在虛無空間中取其上確界(supremum)，此上確界可在極端值，例如$\alpha_X = 0 \cap \beta_M \gg 0$下達到，但在實際應用時，大部份的假說檢定皆落在$\alpha_X = 0 \cap \beta_M = 0$附近。爲了在單一研究（例如全基因體分析）中校正複合型虛無假說造成的p值保守性，我們必須知道$H_0^{(1)}$、$H_0^{(2)}$及$H_0^{(3)}$的比例，然而其比例在資料中的估計不易且不可靠。筆者推導出一個毋需估計三種虛無假說比例的p值估計式(Huang, 2018+a)：

$$F\left(\frac{T_\alpha T_\beta}{\sqrt{\mathrm{Var}(T_\alpha)}}\right) + F\left(\frac{T_\alpha T_\beta}{\sqrt{\mathrm{Var}(T_\beta)}}\right) - F(T_\alpha T_\beta),$$

其中$F(z) = \pi^{-1}K_0(z)$，而$K_0(z)$是 modified Bessel function of the second kind with order zero 且在 R 有可用的函數：besselK (, mu = 0)；而如果資料中大部份的檢定是在虛無假說下，$\mathrm{Var}(T_\alpha)$及$\mathrm{Var}(T_\beta)$皆可由資料中估算。筆者同時也將此一複合型虛無假說的校正方法推廣至縱向的多變量中介檢定（Huang, 2018+b）。

三、結語

因果中介模型爲統計學的新興領域，有許多尚待解決的問題，其理論架構及相應的方法學仍亟需結合其它統計理論進一步發展，例如高維統計、模型選取、存活分析等。另一方面，中介模型也值得被廣泛的應用，它除了能了解因果關係的中介機制，也能作爲整合不同資料的模型架構。

參考文獻

[1] Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. Excerpts reprinted (1990) in English (D. Dabrowska and T. Speed, translators) in *Statistical Science* 5, 463-472.

[2] Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66, 688-701.

[3] Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling* 7, 1393-1512.

[4] Spirites, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

[5] Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika* 82, 669-710.

[6] Baron, R.M. and Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51, 1173-1182.

[7] Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3, 143-155.

[8] Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*, 411-

420. Morgan Kaufmann, San Francisco.

[9] VanderWeele, T. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface* (Special Issue on Mental Health and Social Behavioral Science) 2, 457-468.

[10] Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* 25, 51-71.

[11] Huang, Y.T. and Pan, W.C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* 72, 402-413.

[12] Huang, Y.T. and Cai, T. (2016). Mediation analysis for survival data using semiparametric probit models. *Biometrics* 72, 563-674.

[13] Huang, Y.T. and Yang, H.I. (2017). Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology* 28, 370-378.

[14] Cho, S.H. and Huang, Y.T. (2018+). Causal mediation analysis with sequentially ordered mediators using Cox proportional hazards model. *Statistics in Medicine*, in press

[15] Shih, S., Huang, Y.T. and Yang, H.I. (2018). A multiple mediator approach to quantify the effects of ADH1B and ALDH2 genes on hepatocellular carcinoma risk. *Genetic Epidemiology* 42, 394-404.

[16] Berger, R.L. and Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Sciences* 11, 283-319.

[17] Huang, Y.T. (2018). Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. *Annals of Applied Statistics* 12, 1535-1557.

[18] Huang, Y.T. (2018+a) Genome-wide analysis of sparse mediation effects under a composite null hypothesis. *Annals of Applied Statistics*, in press

[19] Huang, Y.T. (2018+b). Variance component tests of multivariate mediation effects under composite null hypothesis. under revision