

第一章 導論

第一節 研究動機與目的

預測選舉是個普遍存在的興趣，不論基於好奇心的滿足或實用的考慮，選舉預測都是件極具挑戰性的工作（劉義周，1996a：107）。在國內的選舉行為調查研究中，選舉預測也一直是個很重要的主題。有學者認為，選舉預測的研究最足以顯現政治學的科學性，因為預測結果準不準，投票當天晚上就馬上可以映證出來（陳義彥，2003：3）。

根據洪永泰的歸納整理，他認為選舉預測的方法因使用「工具」的不同而有四個方向（洪永泰，1994：93-94）：

1. 利用總體性的資料（aggregate data）從趨勢的角度作籠統的預估。譬如使用各項人口、財經、建設資料來評估選民的結構與意識型態，或使用過去的投票數據來預測未來。不過通常這種方式只能做大方向的評估而不能做單一選區的精確數字預估，大多數的情況都是學者們以此作為事後分析，解釋選舉結果之用。
2. 專家判斷。最瞭解選舉狀況且最具專家資格的莫過於從事競選活動者本人及他們的幕僚。但可惜的是通常在需要選舉預測的時候他們都是當事者，無法也沒有辦法以客觀身份提供這項資訊
3. 從候選人的屬性方面來預測選舉結果。候選人的黨派、家世、學經歷、族群、性別、年齡、財力等都可能被拿來當作預測工具，只是這種方法通常只能預測輸贏，而沒有得票百分比。

4. 使用民意調查結果，提出精確數字的預測。這是到目前為止做得最多、也被認為是最「正統」的預測方式。

不過，儘管使用民意調查來進行選舉預測在目前被認為是最「正統」的預測方式，但還是有一些問題值得我們注意。范凌嘉（2001：26-31）認為使用民意調查進行選舉預測的誤差來源有三大類，分別是調查過程產生的誤差、測量方式與指標的移植、無反應選民。對於選舉預測的研究者而言，調查過程所產生的誤差，在抽樣方法與調查技術沒有提昇之前，是很難獲得改善；而若要改善測量方式的誤差，也有待於政治學理論的逐步充實，因此如果只就研究者所有的資料來改善，其實只能針對無反應的誤差來進行控制。

的確，在許多針對選民投票意向所進行的民意調查中，不管是媒體或是學術單位所公布的資料，都可以看到非常高比例的受訪者不願意回答他們的投票意向，如何對這些未表態的人進行歸類，就是選舉預測的最大挑戰。

在美國，對於未表態選民的歸類，有四種常用的方式，一是將未表態的選民依已表態者的比例，分配給各候選人；二是將未表態者平均分配給兩大主要政黨的候選人；三是如果有在位者參選的話，將未表態者歸給挑戰者；四是使用不需要考慮未表態者的方法，來預測選舉結果（Mitofsky, 1998：233）。而將未表態者依照已表態的比例來分配給所有候選人，在經驗上是最接近選舉結果的（Crespi, 1988：22）。

不過，將未表態者依照已表態者的比例來分配，必須要假定未表態者與表態者基本上是同一群人，所以表現出來的投票行為是一樣的，但是這樣的假設恐怕不容易成立，特別是在台灣，因為歷經威權統治、白色恐怖，有許多受訪者不願意表達他們的投票意向，又或者可能因為生性較為保守，所以比較不敢表達意見。因此，在台灣要利用民意調查進行選舉預測時，必須對民意調查中的未表態選民，投注更多的關切（陳俊明、劉念夏，1998：71）。

筆者以爲，一個好的選舉預測模型，不只要能夠預測的準，還必須能夠適當的解釋選民的投票行爲，因爲從邏輯的角度來看，預測就是解釋的反向過程，如果我們可以瞭解選民在進行投票抉擇時，受到哪些因素的影響，那麼即使受訪者不願意表態，我們還是可以依據他在這些影響因素上所呈現的態度來加以預測其投票對象。在過去許多投票行爲的研究中，對於影響選民投票抉擇因素的研究已經累積了不少的成果，如果可以在這些研究的基礎之上，建立選民投票行爲模型之後，再用以預測未表態選民的投票對象，應該是個相當可行的方式。事實上，過去也有一些研究者運用這種方式來進行選舉預測（張紘炬、林顯毓，1995；張紘炬、丁台怡，2000；張紘炬、黃男璋2000），但大多是僅用已表態者來建立投票模型之後，再根據此模型的參數估計值去推估未表態者的投票意向，這個方法必須假設未表態者之所以不表態是隨機發生的，如此一來，僅以表態樣本所建立的投票模型，才能夠推論其抽樣時所依據的母體。問題是，不願意表態的受訪者極有可能並不是隨機產生的，一旦他們是因爲某些因素而不願意表態時，則不願意表態者與願意表態者是兩群不一樣的人，如此一來，僅以表態樣本所建立的投票模型，其參數估計便會發生偏差，這就是所謂的「選樣偏誤」(selection bias)，以偏差的參數估計去推估未表態受訪者的投票意向，較可能發生錯判的情況。

筆者認爲，在調查訪問中，願意表態的人，在一些政治態度上往往具有比較明顯的傾向，例如具有特定的政黨認同、或是對候選人有明顯的偏好，而不表態的選民則是在這些政治態度上不具有明顯的傾向，僅以這些政治態度較強烈的表態者去建立模型，容易高估模型當中自變數對應變數的影響力，也容易錯估不同政黨認同者或不同候選人偏好與中立無反應選民之間的差距，進而造成模型高估或低估候選人的得票率。

不過，在邏輯上，建立投票模型後，再用以預測未表態受訪者的可能投票對象，的確是一個相當可行的辦法，如果能夠有效的校正「選樣偏誤」的問題，便能夠真正建立一個足以有效推論母體的投票模型。本研究便是希望

經由校正「選樣偏誤」的問題之後，可以得到一個真正足以代表母體的投票模型，再利用這個模型的參數估計值去推估未表態選民的投票意向，以得到更準確的選舉預測結果。

第二節 文獻檢閱

如同之前所說的，預測選舉是個普遍存在的興趣，因此很多民主國家之中都有選舉預測，在澳洲、法國、西德、英國、日本、荷蘭、加拿大、紐西蘭、奧地利及美國等地，自 1949 年至 1986 年之間的 68 次全國性選舉中，就有 155 個選前預測 (Buchanan, 1986)。國內外學界在過去已經累積相當多的研究成果，在建構一個新的選舉預測模型之前，我們應該先回頭去看看過去的研究者曾經建立過哪些模型，他們所依據的理論與方法是什麼，找出他們可資參考及可以改進的地方，並以此為基礎去思考建構一個更好的選舉預測模型。由於國內的投票行為研究大多以美國的研究為基礎，因此以下筆者將先對美國的選舉預測相關文獻進行回顧，接著再針對國內學者所做的相關研究進行探討。

一、美國的選舉預測

從相關的文獻來看，美國的選舉預測大多是針對總統選舉所進行的預測，而且並不全然是政治學家所進行的預測。歷史學家 Lichtman (1996) 檢視自 1860 年以來，圍繞美國總統大選的歷史條件，提出十三項最可能影響總統大選結果的因素，稱之為「通往白宮的十三把鑰匙」(Thirteen Keys to the White House)。這十三個因素主要是去檢驗選民是否會選擇再投票給現任總統的政黨，雖然每一把鑰匙都可以單獨預測選舉結果，但是 Lichtman 設計了一個門

檻(threshold),如果這十三個陳述中,有五個或五個以下的陳述是不正確(false)的,則現任總統的政黨會勝選,若有六個或六個以上的陳述是不正確的,則現任總統的政黨會落敗。這十三把鑰匙的說明如下:

鑰匙一:執政黨的委任(Incumbent-party mandate):期中選舉後,執政黨的眾議員席次比上屆期中選舉後來的多。

鑰匙二:提名競爭(Nomination contest):執政黨的提名過程沒有激烈的競爭。

鑰匙三:現任者(Incumbency):執政黨的候選人是現任總統。

鑰匙四:第三黨(Third party):沒有重要的第三黨或獨立候選人參選。

鑰匙五:短期經濟(Short-term economy):競選活動期間,經濟狀況沒有衰退。

鑰匙六:長期經濟(Long-term economy):總統任期內,年度實際國民平均經濟成長毛額等於或高於前兩任總統任期之平均經濟成長毛額。

鑰匙七:政策轉變(Policy change):在位總統的施政造成國家政策的重大轉變。

鑰匙八:社會動亂(Social unrest):總統任期內無持續的社會動亂。

鑰匙九:醜聞(Scandal):現任政府沒有嚴重的醜聞發生。

鑰匙十:外交或軍事挫敗(Foreign or military failure):現任政府沒有遭到重大的外交或軍事挫敗。

鑰匙十一:外交或軍事成功(Foreign or military success):現任政府有達到重大的外交或軍事成功。

鑰匙十二：現任者魅力（Incumbent charisma）：執政黨候選人有群眾魅力或是位國家英雄。

鑰匙十三：挑戰者魅力（Challenger charisma）：挑戰者沒有群眾魅力或不是位國家英雄。

由於以上這十三把鑰匙是與選民的判斷有關，而不是選舉人團（electoral college），因此所進行的預測是針對選民的投票結果而非選舉人團的選舉結果，在 Lichtman 針對 1860 年到 1992 年的三十四次總統大選所進行的選舉預測中，每一次都可以正確的預測出選民投票結果的獲勝者。¹

雖然這十三個因素看起來似乎簡單易懂，而且涵蓋層面相當廣，包括政治因素（鑰匙一至四）、執政黨的表現和政績（鑰匙五至十一）、及候選人的特質（鑰匙十二至十三）。不過由於必須仰賴主觀的判斷，所以預測結果可能因人而異；再者，Lichtman 對於特定鑰匙的判斷標準有時前後不一，較難讓人信服（黃偉峰，1996：99-101）。最後，這十三把鑰匙的預測效果最多只能達到猜對當選人，而無法較準確的去預測候選人的得票率，以現在的標準來說，是不太足夠的。

爲了能從猜對當選人進步到準確預測得票率，開始有學者進行更科學化的研究。Fair（1988）嘗試以總體資料來建立迴歸模型，模型的應變數是民主黨候選人佔兩黨候選人總得票數之比例，自變數則包括選前兩年的通貨膨脹率、大選年第二季至第三季平均國民生產毛額增加率、民主黨是否爲執政黨、民主黨候選人是否爲現任總統、時間點、選前兩年通貨膨脹率與民主黨是否爲執政黨的交互作用項、以及大選年第二季至第三季平均國民生產毛額增加率與民主黨是否爲執政黨的交互作用項。根據這個模型的預測得票率，在 1916 至 1984 年的十八次選舉中，只猜錯三次當選者，不過由於此模型採用的經濟指標只能在大選後取得，因此也降低了它的預測功能。

¹ 在這三十四次選舉中，有兩次選舉的選民投票結果與選舉人團投票結果不同，分別是 1876 年與 1888 年。

Lewis-Beck 與 Rice (1992) 也同樣採取以總體資料建立迴歸模型的方式進行預測。不過他們所採用的變數與 Fair 有所不同，他們先以總統聲望及經濟情況作為自變數，以 1948 年到 1988 年間，執政黨佔選舉人團票數之比例為應變數，所建立的模型在十一次選舉中，對於當選者的預測雖然僅錯了三次，可是在預測誤差上的表現並不好，平均誤差達到 12.4%，因此 Lewis-Beck 與 Rice 將模型加以改良，加入了政黨實力與候選人吸引力兩個變數，改良後的模型在十一次的選舉中，僅錯一次，而且平均誤差也降低到 5.6%。

以總體資料為基礎的預測模型，雖然可以較為準確的預測得票率，但是模型的建立與檢證都是使用同一筆資料，將會造成「循環論證」的問題（黃偉峰，1996：112）。再者，這些模型並不處理投票者、投票行為以及選舉結果這些面向，對於政治學理論的擴充與精實，幫助相當有限（范凌嘉，1999：11）。

當然除了總體資料之外，美國也有許多學者是以個體資料來進行選舉預測的研究。Kelley 與 Mirer (1974) 認為選民的投票行為有一個簡單的規則，就是先考慮對候選人的偏好，如果對候選人的偏好不明確，再考慮對政黨的偏好。因此他們以密西根大學選舉調查中的八個開放性的題目作為分析的基礎。其中四個題目問受訪者對共和黨及民主黨喜歡和不喜歡的事項，另外四個題目則詢問受訪者對兩個主要政黨提名的總統候選人喜歡及不喜歡的事項，將喜歡的項目減去不喜歡的項目即得到對兩個候選人及對兩個政黨的「淨喜歡值」。在考慮投票對象時，受訪者會先比較兩個候選人的淨喜歡值，如果能分出大小，則會投票給他較喜歡的候選人，如果無法分出高低，再比較對兩個政黨的淨喜歡值，會投票給他比較喜歡的那個政黨，如果連政黨的喜歡程度也分不出高下的話，那麼受訪者可能選擇不去投票、任意投票或投廢票。研究結果發現，依照這個簡單的投票規則，正確預測率在八成到九成之間，可說是個不錯的預測模式。

雖然 Kelley 與 Mirer 的研究與選民的投票行為之間有比以往更深的連結，也提供了一個簡單的投票規則供大家參考，不過在整個模型的設計上，還是有些改進的空間。首先，雖然成功的預測模型是愈簡單愈見其功效（劉義周，1996a：109），但是這個投票規則可能過於簡單了，除了候選人與政黨之外，也許還有其他影響投票的因素存在，其次，僅以受訪者回答的事項數目為基礎進行加減計算，且給予每個事項一樣的權重，如此只考慮量而不考慮質的指標建立，可能有效度上的問題。

相對於 Kelley 與 Mirer 對選民投票行為規則的簡單假設，Markus 與 Converse（1979）所提出的模型就複雜許多。他們認為選民在進行投票抉擇的過程，是一個動態的認知過程，不但現在的因素會影響選民的思考，就連以前的經驗與認知也會影響到他的投票抉擇。因此他們所設計的投票模型相當複雜，考慮的變數包括前一次時間點的投票、政黨認同、議題取向，目前時間點的政黨認同、議題取向、候選人取向，這些變數不但會交互影響，並且會形成對議題得失與候選人特質的看法，然後這兩個變數與政黨認同再進一步產生對候選人的評價，最後影響投票決定。²研究者以 1972 年與 1976 年使用固定樣本連續訪問法（panel study）所得到的資料進行分析，得到對他們的投票假說有力的支持。雖然 Markus 與 Converse 的投票模型並不是為了選舉預測所建構的，也沒有對於候選人的得票率進行實際的預測，不過他們還是自認為有不錯的預測能力，以 1976 年來說，投票模型有九成左右的正確預測率。

Page 與 Jones（1979）同樣以傳統密西根模式主要變項（政黨、候選人評價、政見立場）來進行研究，不過和 Markus 與 Converse 不同的是，他們使用的是單次的調查研究資料。他們的研究發現，在這三個主要變項當中，選民對候選人的評價會強烈影響他對候選人政見立場的認知與感覺，並進而影響投票對象；而政黨認同反而對投票行為的影響並不大。

² 模型中各變數彼此之間的關係相當複雜，請見 Markus 與 Converse（1979：1059）。

以上這兩篇文章雖然都不是作選舉預測方面的研究，但是提供了較新的投票模型，對於影響選民投票行為變數之間的關係與變化，也都做了深入的研究，讓大家瞭解到選民在進行投票抉擇時，其實是一個相當複雜的過程。對於進行選舉預測的研究者來說，這兩篇文章提供了建構模型的思考方向，有其值得參考的地方。

二、台灣的選舉預測

在台灣的選舉行為調查研究中，選舉預測可說是一個很重要的主題，在過去的這十幾年間，國內的學者也發展出許多選舉預測的模型，接下來筆者將根據這些模型所使用的不同方法，一一加以介紹：

（一）以總體資料為輔助工具

以總體資料為輔助工具的選舉預測模型是由洪永泰所首創，他混合使用選區內的民意調查和總體資料，以民意調查為主，總體資料為輔，希望發揮兩者之長，互補其短，稱之為 ADAM 模型（Aggregate Data Assisted Model）（洪永泰，1994：93-110）。

ADAM 模型進行的步驟是先整理選區內歷次選舉投票所的投票紀錄，再挑選指標進行集群分析（cluster analysis），在這個階段所使用的指標指的是與選區選舉特色有關的變數，在每個選區可能有所不同，例如各政黨在該選區過去的選舉得票率、候選人過去的選舉得票率等等，接著對每一種集群結果進行變異數分析與區辨函數判斷法，求得最佳集群組合，也就是選區的政治版圖。之後，將民意調查資料（受訪者可能的投票對象）與政治版圖（受訪者所在的集群所屬）進行交叉分析，然後把每一個集群內未表態的選民人數依照該集群的歷次選舉資料予以研判後分配到各候選人、不去投票或投廢票的格子裡。最後將每一個候選人在每一個政治集群的支持人數相加，換算為

有效票數的百分比，即為得票率的預測。研究結果發現，在對 1993 年台南市長選舉與屏東縣長選舉所進行的選舉預測中，都正確的預測到當選者，在預測得票率與實際得票率的比較上，台南市較差，差距約為 3.1%；屏東市則相當接近，僅差距不到 0.3%。

莊文忠在 2000 年針對總統選舉所進行的選舉預測研究，使用與 ADAM 模型相近的方法，他也是結合總體資料與民意調查資料，來預測總統選舉各組候選人的得票率，與 ADAM 模型的差異在於，並未挑選指標進行集群分析，而是直接以各縣市為分層。他先以近幾次行政首長選舉中，各政黨在各縣市的實際得票率，來估計各政黨在各縣市的實力比例，再將民意調查中各縣市的未表態受訪者人數，依照政黨實力比例予以分配（莊文忠，2000：55-90）。該研究在選前共進行了四波的調查訪問，平均來說，三位主要候選人的預測得票率與實際得票率的差距，約在 7 到 9 個百分點之間，除了得票率預測不太準確之外，在這四波訪問的預測當選對象方面，也僅有一次是正確預測到陳水扁會當選。

造成預測不準確的原因，研究者本身認為可能是未表態受訪者的推估不正確，以及發生策略性投票，因此他改變預測的方法，將連宋佔國民黨實力的比例重新分配，並考慮策略性投票發生的可能性。在問卷當中有一題是關於棄保效應的題目，也就是詢問受訪者當他所支持的第一人選沒有勝選的機會時，會不會改為支持其他候選人，根據受訪者的回答，去計算某一候選人因當選機會小而被放棄時，各組可能獲得的得票率。³研究結果發現，在「棄連效應」下的各組候選人預測得票率，相當接近實際的得票率，誤差不超過 2%。

以總體資料來推算缺失的個人資料，基本上假定個人受到所處環境的強

³ 計算方式為：

被棄者得票率=被棄者原獲支持率*(不會改投者比例+無反應者比例)

被保者得票率=被保者原獲支持率+(被棄者原獲支持率*被棄者遭選民改投被保者的比例)

烈影響，所以從受訪者所居住地區的投票資料，可以推知其投票傾向。同時也假定與選舉有關的環境變項維持穩定，所以我們才能用前次或前幾次的選舉結果來推測這次的可能趨勢。如果這兩項假定同時成立，則這個模式無疑有理論上的優越性（劉義周，1996a：113）。但是一旦這兩項假定其中之一不成立時，則可能產生預測不準確的情況。以 2000 年總統選舉為例，宋楚瑜、許信良脫黨參選，就是違反了第二項假定，究竟脫黨參選者能爭取到多少原來所屬政黨傳統實力的選票，只能參考各政黨支持者會投給各候選人的比例與個人之主觀經驗（莊文忠，2000），如此一來很可能發生誤判，進而導致錯誤推估未表態受訪者的投票傾向。此外，近年來我國發生政黨重組，以過去的投票結果預測未表態者的投票傾向，更可能發生錯判的情況。

而莊文忠在 ADAM 模型之外還考慮到策略性投票，在「棄連效應」發生的前提下，雖然有不錯的預測結果，但是並沒有解決上述 ADAM 模型可能產生的問題。此外，在選舉競爭局勢不明朗的情況下，各種策略性投票的行為可能同時發生，研究者要如何確定某種策略性投票必定發生，是相當大的挑戰。

（二）候選人形象預測模型

梁世武認為在「政黨取向」、「候選人取向」、「政見取向」的投票行為中，以「候選人取向」最具影響力及解釋力，因此他建立了「候選人形象預測模型」(Candidate's Image Predictive Model)，而所謂的候選人形象是選民對候選人特質、政見、立場所有的一種態度或認知（梁世武，1994：108）。

候選人形象指標一共有三個，分別是「一般人支持強度」、「擁護者支持強度」及「最高分法可能得票率」。研究發現，在 1994 年台北市長選舉中，以最高分法所測得之可能得票率，與實際得票率之間非常接近，誤差不到 1%。

雖然候選人形象預測模型有極佳的預測力，不過對於形象因素與政黨認

同、議題取向等因素之間的錯綜複雜關係，缺乏有效的連結與處理（范凌嘉，1999：17）；此外，研究者在模型中以候選人形象為唯一之解釋變項，在極接近選舉日時，或許可以作接近投票意願之預測，因為其他因素如政黨、政見、競選活動效果等的影響到了選戰的後期，可能都被候選人形象所吸收，但是這樣的模型沒有包含其他的變項，實在難以解釋投票行為的主要現象，也就是選民作選擇的法則（劉義周，1996a：112）。

（三）以假設的投票行為法則進行推估

在這部份所要回顧的選舉預測模型，其共通之處在於，研究者假定選民在選擇其投票對象時有一套固定的行為法則，只要將未表態的受訪者透過行為法則進行歸類，即可推估其可能投票對象。

前面提到 Kelley 與 Mirer（1974）所建立的模型也是屬於此類，而為了適用於台灣的環境與資料，劉義周（1996a）針對 Kelley 與 Mirer 的模型進行修正，在判斷標準上，除了候選人與政黨之外，還加上了政見立場，此外他採信表態受訪者所提供的投票意向，對於未表態的受訪者並不預先設定是「先選人後選黨」，而是由受訪者自行決定在政黨、候選人、政見三者之中，何者最為重要，由受訪者的選擇決定他們的投票法則。如果在這三者中，受訪者並未提供優先順序，則維持 Kelley 與 Mirer 所排的順序，依序是候選人、政黨、政見。至於在這三個變數都無法做出判斷的受訪者，他們可能是不投票、依照別的標準來投、或者任意投，在模型中予以不計。研究者將此一改良模型與「先選人後選黨」、「先選黨後選人」兩個模型進行比較，研究結果發現，在 1994 年台北市長選舉及省長選舉中，改良模型的預測力不如其他兩個模型，而其他兩個模型在不同的選舉中也有不同的表現。

劉念夏（1996）以 Kelley 與 Mirer（1974）的研究及劉義周（1996a）的研究為基礎，提出另一個投票行為的法則，他認為在 1996 年的總統選舉中，台灣地區選民的投票行為是循「候選人-政見/議題-政黨」此一方向來思考。進行

的步驟是，先將全體受訪者依照「候選人最高形象評價」、「選民預期候選人解決問題的能力」、「選民的政黨支持」分成十二個選民區隔，⁴然後依據全體樣本在這十二個區隔中，調查資料所實際顯示對候選人的支持度百分比高低，來判斷未表態受訪者的投票傾向。該研究和 Kelley 與 Mirer (1974) 的模型及劉義周 (1996a) 之研究的最大不同在於，對於未表態受訪者的推估並不是直接以受訪者在哪一個區隔就歸到所對應的候選人，而是將未表態受訪者歸類到該區隔中，獲得最高支持百分比，且百分比超過全體受訪者支持百分比的候選人。而當受訪者在三個變數都無法做出判斷時，其推估方式除了根據上述的規則以外，如果獲得最高支持百分比的候選人其支持度低於全體樣本對此候選人的支持百分比時，則將此區隔的未表態受訪者歸類到支持民進黨的候選人。研究結果發現，選前三次的預測結果與實際投票結果之間的誤差介於 1.15% 至 1.75% 之間，預測力可說是相當的不錯。

盛治仁在 2000 年則提出另一個推估未表態受訪者的方式：情感溫度計。所謂的情感溫度計就是請受訪者在一定的範圍內對候選人給分數，分數越高表示好感越強烈，分數越低表示厭惡程度越高，中間的分數表示沒有什麼感覺。⁵他認為，從投票行為的理論來看，對候選人的情感溫度計分數，反應的不只是單純對候選人特質的看法，同時也包含了受訪者政黨認同及議題立場的影響。研究者假定如果受訪者能夠在對候選人的情感溫度計分數上分出高下的話，則會投票給分數較高者 (盛治仁，2000：88-89)。

其執行方式是，先請受訪者對主要的三位參選人打分數，⁶接下來把受訪者對三位主要參選人 (連戰、陳水扁、宋楚瑜) 的情感溫度計分數依照得分

⁴ 區隔 1-4 是對四位候選人有進行評分且給定某位候選人為最高分，區隔 5-8 是無法判斷候選人最高形象評價，但能夠判斷出哪一位候選人的解決問題能力最高，區隔 9-11 是無法對任何一位候選人評出最高分，也無法判斷哪一位候選人解決問題能力最高，但是可以明確表示支持政黨的受訪者，區隔 12 則是在這三項中都無法做出判斷的受訪者。

⁵ 在密西根大學的國家選舉研究計畫 (National Election Study) 中，評分的範圍是 0 到 100 分；而盛文中則是以 0 到 10 分為評分範圍。

⁶ 問卷題目為：「我們想請您為幾個政治人物打分數。從 0 到 10 分。非常喜歡是十分，非常不喜歡是 0 分，普通沒感覺是五分。請問您給 XXX 幾分？」。

高低作比較排列，將未表態受訪者分爲七類，分別是最喜歡連、最喜歡扁、最喜歡宋、較喜歡連宋、較喜歡連扁、較喜歡宋扁、給連扁宋一樣的分數。研究者將最喜歡某位候選人的該組樣本數歸類到該候選人的支持者之中，再將偏向兩位候選人的受訪者平均分配給相對應的兩位候選人，至於無法在三位候選人當中分出喜歡程度高低的受訪者，則予以扣除。最後的預測結果與實際得票結果比較，亦有不錯的預測力。

不過研究者本身也承認，使用情感溫度計雖然可以有較高的預測能力，但是在理論方面的解釋力是相對較薄弱的，因爲它本身已經包含了政黨、候選人和議題立場的影響（盛治仁，2000：94）。

綜合以上三種模型來看，雖然研究者對於未表態選民都提出了一套投票行爲規則的假設，不過除了劉義周（1996a）的研究之外，都只針對一次選舉進行預測，這些假定的投票模式是否可以成立，必須要多經過幾次選舉的檢驗。筆者以爲，或許選民真有其投票行爲規則，但是不見得在每一種類型的選舉中，所採取的模式都是相同的，在劉義周（1996a）的研究中，我們就可以發現，同一種模型在不同的選舉當中，有不同的表現。因此從事不同的選舉之預測，要先從各該選舉的外部條件來判斷與使用適當的模式，各模式中所包含的變項，可能不太一樣（劉義周，1996a：125）。

（四）統計方法與政治學理論的結合

陳義彥（1994）利用「政黨因素」、「候選人形象」、「統獨政見」及「族群因素」對受訪者進行集群分析（cluster analysis）。再將所得到的四個集群與表態受訪者進行交叉分析，由此判斷各集群的可能投票傾向，然後以此推論各集群中未表態受訪者的投票傾向。不過由於該文是屬於回溯性質的研究，並未進行具體的預測，因此尚無法斷定其預測力的好壞。

另一個很常被運用在選舉預測的統計方法則是「對數成敗比模型」（logit

model)。張紘炬、林顯毓（1995）以受訪者的性別、年齡、籍貫、教育程度及政黨支持作為自變數，受訪者的投票對象為應變數，建立對數成敗比模型，依照模型所估計出的參數去計算未表態選民投票給各候選人的機率。研究發現，此模型對於 1994 年台北市長選舉有不錯的預測力，不過有學者認為，單以選民的人口學變項再加上一個類似於政黨認同（但仍不是政黨認同）的類別變項，有此解釋力，與選舉行為研究的主流理論有不小的差異，其可適用性還待觀察（劉義周，1996a：112）。

延續著相同的方法，張紘炬、丁台怡（2000）及張紘炬、黃男瑋（2000）分別針對 1998 年台北市長選舉、高雄市長選舉進行選舉預測研究。台北市長選舉的對數成敗比模型，是以性別、年齡、教育程度、籍貫為自變數；高雄市的自變數則多加入政黨支持變數。在這兩項研究中，除了對數成敗比模型之外，還運用區辨分析（discriminant analysis）來進行預測，研究者並比較兩者的預測力。研究結果發現，在台北市長選舉的預測上，兩種統計分析方法都有不錯的預測力，不過區辨分析略優於對數成敗比模型；在高雄市長選舉方面，對數成敗比模型的表現則是比區辨分析差了許多。

在上述兩項研究中，區辨分析的預測都優於對數成敗比模型，因此區辨分析或許也是進行選舉預測的可行之路。此外，筆者以為，對數成敗比模型的好壞，與被挑選進入模型當中的自變數有極大的關係，僅放入人口學變項與政黨變數，並不足以解釋選民複雜的投票行為，這恐怕是造成對數成敗比模型預測力不佳的最大原因。

盛杏媛（1998）同樣以建立投票模型的方式來進行選舉預測，但與過去研究不同的是，他認為根據受訪者選前所回答的投票意向來進行選舉預測會面臨三個問題，一是用表態者推估未表態者可能產生偏誤、二是受訪者可能沒有提供真確的答案、三是受訪者可能會改變投票意向，要解決這三個問題必須認識到選民投票不確定性與測量誤差的可能性，也就是說選民投票給某位

候選人的機率不是二分法的 1 或 0，而是介於 1 與 0 之間。

預測進行的方式是，以「候選人評價」、「政黨認同」、「對政府的評價」及「省籍」為自變數，「受訪者投票對象」為應變數，建構選民投票抉擇的「多元對數成敗比模型」(multinomial logit model)，然後估算每個受訪者投給每一位候選人或不願意表態的機率。研究發現，選民如果投給某一個候選人的機率高，則他投票給該候選人的可能性就大，且不易變更，而投票給任一候選人的機率都不大的選民，他的投票較有可能改變。由於選民不確定性及測量誤差的存在，因此研究者在進行選舉預測時，不以受訪者選前所表達的投票意向為根據，而是根據投票模型所估算的機率來判斷選民的投票對象，如果受訪者投給某位候選人的機率大於投給其他候選人或不表態的機率，則判斷他會投給該位候選人，如果受訪者不表態的機率大於投給其他候選人的機率，則無法預期投票人選，扣除無法預期投票人選的比例，可以估算出候選人的相對實力，也就是候選人的預測得票率。此一估測方式對 1994 年台北市長選舉及 1997 年桃園縣長補選的預測還算準確，候選人預測得票率與實際得票率的差距大約在 5 個百分點左右，但是對 1994 年台灣省長選舉及 1996 年總統選舉的預測則不那麼準確。

范凌嘉(1999)針對台灣地區八個縣市的縣市長選舉所建立的預測模型，則是以「對數迴歸模型」(logistic regression model)結合社會學與人口學變項、社會心理學變項、環境因素來進行選舉預測。在計算候選人的可能得票率時，他採取了一個新的方式，在傳統對數迴歸模型中，對於受訪者選票的預測是歸給機率值最大的候選人，范凌嘉認為這樣的方式並未考慮到選民的不確定性，因為已表態的受訪者本身已具不確定性，而未表態的受訪者，其投票行為較已表態受訪者有更高的不確定性，研究者若僅單純地比較各候選人的機率值大小，便要強塞一個候選人給這位受訪者作為其投票抉擇預測，等於是犧牲對於不確定性的關照(范凌嘉，1999：33)。因此他將每位候選人在受訪者所得到的機率值，處理之後予以累加，來作為預測值。研究結果發現，在

各縣市的預測誤差均不超過抽樣誤差，預測力相當的好。

（五）其他模型

劉文卿（1995）運用遺傳演算法的觀念，建立基因模型。其基本假設為，每一位受訪者均有一選舉染色體，該染色體決定了受訪者的投票行為，受訪者的選舉染色體利用問卷建立，並依據受訪者的染色體建立候選人的選舉染色體，選民會投票給與其染色體最類似的候選人。針對 1994 年台北市長選舉的研究結果，預測平均誤差在 2% 上下。不過，基因模型最大的問題在於其缺乏政治學理論的支撐，讓我們只能知其然而不知其所以然。

李錦河與溫敏杰將行銷學中「產品是各種屬性的集合，產品屬性決定顧客對產品滿意程度」之觀念運用至選舉預測上，他們認為候選人的條件（特質）與產品屬性具同樣意義，若愈能符合選民（市場）的需求，則當選的機率就愈高（李錦河、溫敏杰，1998：21）。執行的方法是，先利用根據候選人屬性條件而設的五個題目建立「選民需求指標」，然後再依據各個候選人票源流出與流入的狀況，加以調整及修正而得到實際預估得票率。研究結果顯示，預測力有不錯的表現。

此外，純粹從統計學的角度來進行選舉預測的是溫敏杰、杜宜軒、李錦河（2000）的研究，他們運用大數法則與不偏估計法來進行選舉預測。也就是說，每一次的獨立抽樣調查時，都可以根據表態的受訪者而得到某位候選人的支持率，在累積多次調查之後，將這些支持率相加取其平均數之後，會接近母體的真实比例。由於研究者在計算候選人支持率時，僅計算表態者的態度，因此他們的假設是，未表態選民與表態選民有相同的支持率。研究結果顯示，1994 年台北市長選舉的預測結果誤差小於 1%，1998 年台北市長選舉的預測結果誤差約為 3.5%，有不錯的預測力。緊接著，溫敏杰與李錦河運用相同的方法，對 1998 年台南市立法委員選舉進行選舉預測，平均誤差為 3.1% 左右，結果也是相當不錯（溫敏杰、李錦河，2000）。

徐永明與林昌平（2003）同樣也是使用多次調查的資料，以時序模型進行選舉預測，他們認為在進行選舉預測時，若只採用單筆的資料進行分析，可能無法包含選舉期間各重大議題的影響性，並且在面對到樣本中的各個次群體時，常會碰到樣本數過少的問題，因此他們將單一性的民調分析加入時間變項的考量後進行跨期的時間序列分析，如此民調預測將包含各次重大議題的影響性，且可以將調查分析的層次拓展至時間的面向，解決樣本數過少的問題。而為了釐清在兩次民調之間的變化究竟是真實的變化或只是隨機樣下的抽樣誤差，徐永明與林昌平使用 Kalman filtering and smoothing algorithm（卡爾曼濾子與平滑估計值）⁷來進行估計。他們以 TVBS 在 2000 年總統大選前所做的一系列民調資訊為基礎，計算出一個卡爾曼平滑估計值，這個估計值雖然低估陳水扁與宋楚瑜的得票率，高估連戰的得票率，不過仍準確的預測連戰、宋楚瑜及陳水扁的排名順序，而且比 TVBS 的預測結果準確。

在多數使用個體資料進行選舉預測的研究之外，盛治仁（2003）嘗試完全使用總體資料來進行選舉預測。作者觀察過去 15 年來台灣的各项重要選舉，發現民進黨的得票率變化並沒有太大幅度的波動，因此他嘗試以民進黨過去的得票相關模式來建立模型，作為預測 2004 年總統選舉結果的根據。研究進行的程序為，先以民進黨 1997 年縣市長（加上 1998 年北高市長）及 1998 年立法委員選舉的地區得票率為自變數，以民進黨 2000 年總統選舉地區得票率為應變數，建立一個迴歸模型，利用這個模型的迴歸係數，套入 2001 年立法委員和縣市長選舉中，民進黨在各地區的得票率，來得到 2004 年總統選舉民進黨候選人的地區得票率。在該研究中，作者分別以縣市及鄉鎮為分析單位，建構兩個模型，其中以縣市為分析單位的模型，預測結果為民進黨候選人可得到 49.63% 的選票；以鄉鎮市為分析單位的模型，則預測民進黨候選人得票率為 43.03%。與 2004 年總統選舉的結果相比，以鄉鎮市為分析單位的模型，預測力較差，以縣市為分析單位的模型，則表現出不錯的預測力。

⁷ 關於 Kalman filtering and smoothing algorithm（卡爾曼濾子與平滑估計值）的計算方式，請參閱徐永明與林昌平（2003），在此不多做說明。

綜合以上對於國內外選舉預測文獻的探討，我們可以發現，幾乎每個選舉預測模型都有不錯的預測力，但是多數的研究僅以一次選舉為預測對象，其預測模型是否穩定還有待觀察。

此外，有的模型完全站在統計學的角度進行，缺乏與政治學相關理論的結合，讓人知其然而不知其所以然；有的模型雖然是根據政治學相關理論而建構，但是僅以一兩個變數來代表所有可能影響選民投票的因素，在解釋力上稍嫌薄弱。

在這些預測模型之中，ADAM 模型以及將統計方法與政治學理論結合建立模型這兩種方式，可說是較能夠兼具預測力與解釋力的預測方法。但是，誠如之前所提到的，ADAM 模型基本上假定個人受到所處環境的強烈影響，所以從受訪者所居住地區的投票，可以推知其投票傾向。同時也假定與選舉有關的環境變項維持穩定，所以我們才能用前次或前幾次的選舉結果來推測這次的可能趨勢。以過去台灣的政治情勢來看，與選舉有關的環境變項並不是很穩定，比如說候選人常有脫黨參選的情況發生，也造成 ADAM 模型在某些選舉上的預測上並不是很準確。

將統計方法與政治學理論結合則是另一個較為可行的方法，但是在過去的研究中，多數是將已表態選民與未表態選民視為同一類，但他們很有可能並不是同一類，而這也正是為何要特別注意「選樣偏誤」的主要原因，在下面的章節中，筆者將介紹如何處理「選樣偏誤」的問題。

第三節 選樣偏誤在政治學上的應用

目前在政治學各個領域中，有許多學者都是以調查訪問所蒐集到的資料來進行研究，運用科學方法抽出可以代表母體的樣本，再根據樣本的特性

來推論母體。這個研究方法可以說是目前政治學界相當具有科學性的研究方法之一，不過有一點值得注意的是，只要是利用調查訪問所蒐集到的資料來進行分析，就可能遇到選樣偏誤的問題。一般說來，在調查訪問的過程中，研究者無可避免地會遇到無反應的問題，也就是說，受訪者不願意提供他對某個問題的意見、或是他根本沒有意見，因此在每個問題上，我們都可以將受訪者簡單的區分成有反應與無反應兩個子樣本（subsample），如果說無反應這個部分的子樣本是隨機產生的，那麼僅以有反應的子樣本來推論母體，並沒有太大的問題；但是如果無反應的子樣本並不是隨機產生的，那麼僅以有反應的子樣本來推論母體，從較為簡單的樣本平均數、標準差到較為複雜的統計模型的參數估計都有可能發生偏誤。

在許多的投票行為研究中，研究者皆嘗試建立投票抉擇模型，並試圖解釋哪些因素會影響選民的投票行為。大多數的研究者都是以有去投票且清楚說明投票對象的受訪者為建立投票模型所依據的樣本，如此一來，沒有去投票或是不願意回答投票對象的受訪者就被排除在外，如果不去投票或是不願意回答投票對象的受訪者是隨機出現的，那麼我們便不用太擔心選樣偏誤的問題；不過在過去的研究中，我們可以發現不去投票或是不願意表達投票對象的受訪者可能並不是隨機出現的。同樣地，從過去的研究我們也可以發現，在選前願意表態與不願意表態的可能是兩群人（盛治仁，2000：79-80），因此當我們以表態者建立投票抉擇模型時，必須考慮到選樣偏誤的問題，才能夠得到正確的參數估計。以下便針對幾種處理選樣偏誤的方法加以說明。

雖然本文所要建立的選舉預測模型（結果方程式），應變數是投票對象，也就是類別變數，不過筆者將先說明當結果方程式的應變數是連續變數時的處理方式。Heckman 處理選樣偏誤的方式是從建構無反應過程的模型而來的，也就是說既然受訪樣本在某個問題的無反應並不是隨機產生的，那麼如果我們能建構受訪樣本為何願意回答的模型，便有方法可以矯正僅以有反應子樣本所建構的模型的偏差（Brehm, 1993：121）。

Heckman (1979) 使用了兩階段的方法來處理選樣偏誤的問題，首先考慮以下兩個方程式：

$$Y_{1i}^* = \beta_1' X_{1i} + u_{1i}^* \quad (1)$$

$$Y_{2i} = \beta_2' X_{2i} + u_{2i} \quad (2)$$

第一個方程式是選樣方程式 (selection equation)，方程式中的自變數是影響樣本被選入的因素，應變數則是是否被選入，如果 $Y_{1i}^* \geq 0$ ，則樣本才能夠被選入第二個方程式，也就是結果方程式 (outcome equation) 中進行分析，而第二個方程式，由於它的應變數是連續變數，因此使用多元迴歸 (multiple regression) 來進行參數估計，在迴歸模型中，對於誤差項的基本假定是，誤差項的分配是平均數為 0 的常態分配，誤差項與模型中其他的自變數的關係是獨立的，如果此模型是依據全部的樣本所建立，那麼誤差項 (u_{2i}) 將符合這個基本假定。但是實際上能夠進入結果方程式 (outcome equation) 中的樣本並不是全部的樣本，在第一個方程式，也就是選樣方程式 (selection equation) 中，如果 $Y_{1i}^* \geq 0$ ，該樣本才能夠被選入結果方程式 (outcome equation) 中進行分析。如此一來，僅使用被選入樣本所建構的迴歸模型，其誤差項將違反迴歸模型的基本假定，誤差項與模型中其他自變數之間的關係並不是獨立的，因為這個原因，模型的參數估計值會有所偏差。

Heckman 的兩階段校正程序為，第一階段先估計選樣方程式 (selection equation)，也就是方程式 (1)，由於應變數是一個二分變數 (1 是選入、0 是未被選入)，因此他使用 probit model 來進行分析，先利用估計出的參數去計算應變數的預測值，再依據應變數的預測值建立 Inverse Mill's Ratio (IMR)，⁸簡單的說，IMR 是樣本被選入觀察的機率。第二階段則是將 IMR 當作一個新增的自變數放入結果方程式中 (outcome equation) 進行參數估計的矯正，如此一來，原本的誤差項 (u_{2i}) 中與其他自變數相關的部分將被移除，新的誤差項

⁸ $IMR = \phi(\beta_2' X_2) / \Phi(\beta_2' X_2)$ ， ϕ 是常態機率密度函數 (normal probability density function)； Φ 是累積常態機率函數 (cumulative normal probability function)。

將符合迴歸模型的假定，而其他原有的自變數所得到的迴歸係數也會是正確的，此外，IMR 的迴歸係數代表選樣方程式與結果方程式的誤差項之間的相關程度，其數值大小便代表選樣偏誤的嚴重性。

Achen 同樣使用兩階段的估計方式，但是使用的統計方法與 Heckman 有所不同，Achen 認為 Heckman 在第一階段的選樣方程式中使用 probit model 來估計，雖然所得結果較為精確，但是過程過於複雜，因此他建議使用線性機率模型（linear probability model, LPM）來處理（Achen, 1986：100-105），LPM 可說是迴歸模型的一種變形，因為它的應變數不是連續變數，但是仍然將它視為線性模型來進行估計，對於自變數的迴歸係數的解釋方式是，在其他條件不變的情況下，X 每變動一個單位，造成 Y 多少機率的變動。由於應變數不是連續變數，因此會有變異異質性（heteroskedastic）的問題，不過在沒有違反其他最小平方法（ordinary least squares, OLS）基本假定的情況下，運用 OLS 進行參數估計，其迴歸係數仍然是不偏的，但是標準誤是不正確的，必須使用一般化最小平方法（generalized least squares, GLS）來進行校正（Achen, 1986：40-41）。⁹LPM 與 probit 的差別在於對於誤差項的假定不同，probit 假定誤差項的分配是常態分配，而 LPM 則假定誤差項是均一（uniform）分配。Achen 首先在第一個階段利用 LPM 去估計選樣方程式的參數，將所得到的參數去計算應變數的預測值，如果預測值超過 0.99 或低於 0.01 則過錄為這兩個臨界值，接著再將原始的應變數觀察值減去預測值，得到誤差項。之後再將誤差項當作新增加的自變數放入第二階段的結果方程式中，進行迴歸分析，便可以得到校正後的迴歸係數。

⁹ 利用 GLS 校正標準誤的程序為：

- A. 將原有模型先使用 OLS 進行參數估計。
 - B. 接著針對每一個觀察值建立 Y 的預測值 p ，將大於 0.99 的預測值過錄為 0.99，小於 0.01 的預測值過錄為 0.01，令 $q=1-p$ ， $s=\sqrt{pq}$ 。
 - C. 將迴歸模型中的自變數都除以 s ，常數項為 $1/s$ ，對這些重新處理的變數使用 OLS 進行參數估計，則所得到的迴歸係數與標準誤都將是正確的。
- 在程序 C 中，每個自變數都除以 s ，換句話說，也就是將所有自變數都乘上一個 $w=1/s$ 的權值，因此整個校正程序又可以稱為加權最小平方法（weighted least squares, WLS）。

上面兩個方法都可以有效的校正選樣偏誤，但是這兩種方法的結果方程式的應變數都是連續變數，當結果方程式的應變數是類別變數時，誤差項將不是呈現常態分佈，兩個誤差項的聯合分配應呈現常態分配的假定將被違反，因此所估計出的參數將是不一致（inconsistent）的（Brehm, 1993：123）。為了解決這個問題，學者們又發展出其他的方法。

Achen 在 LPM 的基礎之下，發展出另外一個兩階段的估計方式，在第一階段與之前相同，仍然使用 LPM 來估計選樣方程式，並且同樣將原始的應變數觀察值減去預測值，得到誤差項（ \hat{u}_{1i} ）。第二階段則與之前不同，他建構一個非線性的結果方程式，如下所示：

$$Y_{2i} = \beta_2' X_{2i} + \alpha \hat{u}_{1i} (\beta_2' X_{2i}) (1 - \beta_2' X_{2i}) + v_{2i}$$

其中 $\beta_2' X_{2i}$ 就是原本的結果方程式等號的右邊， \hat{u}_{1i} 是第一個階段所求出的誤差項， v_{2i} 則是新的誤差項。運用非線性最小平方法（nonlinear least squares）來估計模型的參數，而 α 的大小就代表選樣偏誤的嚴重性。

Achen 使用上面這個方法，對罪犯保釋（pretrial release）進行研究（Achen, 1986：107-129）。在保釋的過程中，可能產生兩種錯誤，一是無辜者被不必要的拘留，二是具有危險性的罪犯被保釋，因此，對於罪犯交保制度的評估，就是法官是否讓適當的罪犯獲得保釋，而是否適當便是判斷當罪犯被保釋時，可能再度犯罪的機率。為了達到這個目的，研究者必須能夠比較被保釋者的實際行為與無法保釋的罪犯如果被保釋時的預期行為。

過去的研究進行方式是，以被保釋的罪犯是否在審判之前再度犯罪為應變數，以可能影響罪犯再度犯罪的自變數建立模型，得到變數的參數估計值，根據這些參數估計值可以去計算無法保釋者如果被保釋時，可能再度犯罪的機率，問題是，這其中便可能發生選樣偏誤的問題，因為建立模型所根據的資料是被保釋者的資料，但是法官在選擇罪犯是否可以被假釋時，可能是在

一定的法律程序或是某些政治壓力下，讓較為安全或較為可信的罪犯保釋，因此無法保釋的罪犯並不是隨機產生的，在這樣的情形下，應該考慮到選樣偏誤的問題。Achen 的研究結果發現，在加入選樣方程式，考慮選樣偏誤的情形下， α 小於 0，表示如果不考慮選樣偏誤的話，將低估被保釋罪犯的再犯罪率。

另外一個方法則是 Dubin 與 Rivers (1989) 所發展出來的方法。與 Heckman 相同的是，他們也是去模型化選樣方程式與結果方程式，不過他們並不是用兩階段的估計方式，而是採取最大概似法 (Maximum Likelihood Estimation) 將選樣方程式與結果方程式一併進行估計 (Sheng, 1994 : 224)。同樣的，我們先考慮以下兩個方程式：

結果方程式為：

$$Y_{1i}^* = \beta_1' X_{1i} + u_{1i}$$

$$Y_{1i} \text{ 的定義為, } Y_{1i} = \begin{cases} 1, & \text{如果 } Y_{1i}^* > 0 \\ 0, & \text{其他情況} \end{cases}$$

選樣方程式為：

$$Y_{2i}^* = \beta_2' X_{2i} + u_{2i}$$

$$Y_{2i} \text{ 的定義為, } Y_{2i} = \begin{cases} 1, & \text{如果 } Y_{2i}^* > 0 \\ 0, & \text{如果 } Y_{2i}^* \leq 0 \end{cases}$$

由於 Y_{1i} 發生的機率是在 Y_{2i} 發生的情況下的條件機率，我們必須考慮聯合常態分配函數 (joint normal distribution function)。誤差項 u_1 和 u_2 的聯合累積密度函數可以用 $F(u_1, u_2; \rho)$ 來表示，而誤差項個別的邊際分配則是 $H(u_1) = F(u_1, \infty; \rho)$ 與 $H(u_2) = F(-\infty, u_1; \rho)$ 。函數中所列出來的 ρ ，便是 u_1 和 u_2 之間的相關係數。

將這兩個方程式擺在一起思考的話，將有三種可能的結果：第一種是被選入可觀察的樣本，且 Y_{1i} 等於 1 (Y_{1i} 與 Y_{2i} 皆等於 1)，第二種是被選入可觀察的樣本，且 Y_{1i} 等於 0 ($Y_{1i}=0$ 且 $Y_{2i}=1$)，第三種則是未被選入可觀察的樣本 ($Y_{2i}=0$)，接下來便是去計算這三種可能的結果發生的機率。

首先以 $G(\cdot, \cdot; \rho)$ 代表 $F(\cdot, \cdot; \rho)$ 的右尾 (upper tail) 機率，例如：

$$G(u_1, u_2; \rho) = \Pr(u_{1i} > u_1, u_{2i} > u_2) = 1 - H(u_1) - H(u_2) + F(u_1, u_2; \rho)$$

接者計算樣本被選入觀察的機率：

$$Q_i(\beta_2) = \Pr(Y_{2i}=1 \mid X_{1i}, X_{2i}) = \Pr(Y_{2i}^* > 0) = 1 - H(-\beta_2 X_{2i})$$

三種可能結果的機率如下所示：

(一) 被選入可觀察的樣本，且 Y_{1i} 等於 1

$$\begin{aligned} P(\beta_1, \beta_2, \rho) &= \Pr(Y_{1i}=1, Y_{2i}=1 \mid X_{1i}, X_{2i}) \\ &= \Pr(Y_{1i}^* > 0, Y_{2i}^* > 0 \mid X_{1i}, X_{2i}) \\ &= G(-\beta_1 X_{1i}, -\beta_2 X_{2i}) \end{aligned}$$

(二) 被選入可觀察的樣本，且 Y_{1i} 等於 0

$$Q_i(\beta_2) - P(\beta_1, \beta_2, \rho)$$

(三) 未被選入可觀察的樣本

$$1 - Q_i(\beta_2)$$

將三種可能的結果合併之後，可以得到一個對數概似函數 (log likelihood function)：

$$\begin{aligned} L(\beta_1, \beta_2, \rho) &= \sum_{i=1}^n Y_{2i} Y_{1i} \log P_i(\beta_1, \beta_2, \rho) \\ &\quad + \sum_{i=1}^n Y_{2i} (1 - Y_{1i}) \log (Q_i(\beta_2) - P_i(\beta_1, \beta_2, \rho)) \\ &\quad + (1 - Y_{2i}) \log (1 - Q_i(\beta_2)) \end{aligned}$$

最後利用最大概似法去求得 $\theta = (\beta_1, \beta_2, \rho)$ 的估計值。

其中 ρ 的估計值大小，代表了選樣偏誤的嚴重性，估計值的方向所代表的意義是，如果 ρ 大於 0，表示在不校正選樣偏誤的情況下，將高估結果方程式中， $Y=1$ 發生的機率，如果 ρ 小於 0，則表示在不校正選樣偏誤的情況下，將低估結果方程式中， $Y=1$ 發生的機率。

除了注意估計值大小及方向之外，還必須注意到的是，對於 $\rho = 0$ 的統計檢定，筆者前面曾經提到，函數中所列出來的 ρ ，便是 u_1 和 u_2 之間的相關係數，也就是選樣方程式與結果方程式之間是否存在選樣偏誤的問題，在統計上的虛無假設是 $\rho = 0$ ，Dubin 與 Rivers 提出三種檢定的方法，分別是 Wald's method、概似比檢定 (likelihood ratio test, LR)、score test 三種 (Dubin and Rivers, 1989: 378-381)。¹⁰ 本文將採取概似比檢定 (likelihood ratio test, LR) 的方法對 $\rho = 0$ 進行統計檢定：

$$LR = -2 (L(\tilde{\beta}_1, \tilde{\beta}_2, 0) - L(\hat{\beta}_1, \hat{\beta}_2, \hat{\rho}))$$

其中 $L(\tilde{\beta}_1, \tilde{\beta}_2, 0)$ 的計算方式為，先將結果方程式與選樣方程式分別用 probit 進行估計，求到 $\tilde{\beta}_1$ 、 $\tilde{\beta}_2$ ，以及控制在 $\tilde{\beta}_1$ 、 $\tilde{\beta}_2$ 之下的 log likelihood 值，將兩個 log likelihood 相加，即是 $L(\tilde{\beta}_1, \tilde{\beta}_2, 0)$ ；而 $L(\hat{\beta}_1, \hat{\beta}_2, \hat{\rho})$ 就是同時考慮結果方程式與選樣方程式的 log likelihood 值。概似比的分配近似於自由度為 1 的卡方分配 (chi-square distribution)。

Dubin 與 Rivers (1989) 使用上述的方法，運用 1984 年美國國家選舉研究 (American National Electional Study, ANES) 的資料，檢驗在 1984 年美國總統選舉中，只以投票者所建立的投票模型是否存在選樣偏誤的問題。研究結果顯示，的確存在選樣偏誤的問題， ρ 的參數估計值小於 0，表示只以投票者所建立的投票模型低估了選民投票給雷根的機率，換句話說，沒有去投票的選民比去投票的選民更傾向投票給雷根。

¹⁰關於 Wald's method、score test 的詳細說明，請參考 Dubin and Rivers(1989: 378-381)，在此不多加說明。

盛杏媛使用 Dubin 與 Rivers 的方法，針對三種基本的投票模型，分別是社會學模型 (sociological model)、社會心理學模型 (social psychological model) 及 Jacobson 的國會投票模型 (Jacobson's Congressional vote choice model) 進行研究，檢驗在這三種投票模型中是否存在選樣偏誤的問題 (Sheng, 1994)。

研究結果發現，社會學模型在 1988 年美國國會選舉中，存在選樣偏誤的問題，在不處理選樣偏誤的模型中，教育程度、收入與年齡的參數估計值明顯高出處理選樣偏誤後的模型參數；不過在 1988 年美國總統選舉與 1990 年美國國會選舉中，社會學模型並沒有存在明顯的選樣偏誤問題。社會心理學模型在 1988 年總統選舉中，也存在著選樣偏誤的問題，在不處理選樣偏誤的情況下，會高估選民的政黨認同、及選民對政府表現評價這兩個變數對投票的影響力，此外，也會低估議題偏好、候選人評價這兩個變數對投票的影響力。Jacobson 的國會投票模型在 1988 年美國國會選舉及 1990 年美國國會選舉中，都存在著選樣偏誤的問題，在變數的影響力表現上，選民居住選區的現任國會議員黨籍這個變數對於投票的影響力，在兩個模型中有極為明顯的差距。綜合來說，三種投票模型或多或少都存在著選樣偏誤的問題，尤其是社會學模型與 Jacobson 的國會投票模型，而且選樣偏誤的問題會使研究者高估或低估自變數的影響力，因此思考選樣偏誤是有必要的。

