選舉研究
第十二卷第二期（94/11），pp.01-44

選樣偏誤模型在選舉預測上的應用

周應龍*、盛杏湲**

《本文摘要》

在從事選舉預測時，研究者常面臨受訪者不告知其投票對象的問題，若僅以告知投票對象的受訪者作選舉預測，將無可避免地造成選樣偏誤的問題。本文主要目的在於評估選樣偏誤對於投票模型估計所造成的影響，並且試圖藉由矯正選樣偏誤所造成的問題，得到較正確的參數估計值，並進而作更精確的選舉預測。在本文中，我們採取 Dubin 與 Rivers 所發展出來的二變量選樣偏誤模型（bivariate selection bias model）為研究方法，為了檢視選樣偏誤模型在選舉預測上的穩定性，我們將之應用在五次不同的選舉中。結果發現在五次選舉中，未校正選樣偏誤（也就是只以願意回答投票對象者加以預測），都會造成高估自變數對應變數的影響，因為願意回答投票對象者往往是政治偏好較強或較確定的受訪者，也因此會造成選舉預測的偏誤。當我們校正選樣偏誤後，在四次選舉中都發揮了極好的效果，預測的誤差都比原本不校正選樣偏誤來得更小，且誤差都不超過 1.16%，可謂相當地準確。唯有在一次選舉無法發揮校正的效果，但是即便如此，也並不會比不校正更差。我們認為這樣的效果顯示，選樣偏誤模型是一個相當可以信賴的選舉預測工具。

關鍵詞：選舉預測、選樣偏誤模型、調查研究中的無反應

---

* 國立政治大學政治學系博士生。
** 國立政治大學政治學系教授。

壹、前言

選舉預測是民意調查與選舉研究中一個普遍關注的課題，近些年來，每逢選舉，各式各樣的選舉預測競相出爐，不僅佔用大眾媒體相當的篇幅，也是學術研討的重要議題。這些選舉預測不僅提供一般民眾誰勝誰敗、勝敗差距多少的訊息，也是政黨與候選人作爲選舉策略的參考。同時，選舉預測也提供了社會科學研究者一個不可多得的將學術理論及方法與事實加以印證的機會，因爲多數學術研究主題沒有真實的結果，而選舉預測是否準確則可以從最後的選舉結果來加以評判。

然而，以民意調查來作選舉預測勢必面臨到一個困難，許多受訪者在被詢問投票對象時，沒有提供確切的投票對象，尤其是距離選舉日愈遠，不表態的選民就愈多，那麼如何來推估這些不表態選民的投票意向就是選舉預測所面臨的眾多挑戰之一。大多數選舉預測是以表態者去推估不表態者，如果受訪者表態或不表態是隨機產生的，則表態者與不表態者沒有差別，那麼以表態者去推估不表態者，沒有太大的問題。然而，如果受訪者的表態或不表態，是受到某種因素的影響，則表態與否是系統性，而非隨機性的，亦即表態與不表態是兩群不一樣的人，則以表態者去推估不表態者的投票對象必然產生偏差。基於過去有關台灣選民投票行爲的研究，可以發現那些對政治較沒有興趣、沒有固定政黨認同或候選人偏好者、對政治較爲敏感、教育程度較低、年齡較長的選民比較偏向不表態，而這群人極可能與表態者有不同的投票對象，那麼若以表態者來推論未表態者，則產生選樣偏誤（sample selection bias）無可避免。

本文主要目的在於評估選樣偏誤對於投票模型的估計以及選舉預測所造成的影響，並且試圖藉由選樣偏誤模型，來矯正選樣偏誤所造成的問題。在本文中，我們採取Dubin與Rivers（1989）所發展出來的二變量選樣偏誤模型（bivariate selection bias model）爲研究方法。爲了檢視選樣偏誤模型在選舉預測上的穩定度，本研究以五次選舉爲例，包括兩次總統選舉（2000年與2004年）與三次地方首長選舉（2001年台北縣長選舉、2002年台北市長與高雄市長選舉）。這幾次選舉有全國性與地方性的選舉，同時有選舉勝負懸殊（2002年台北市長選舉）與勝負差異不大的選舉（其它四次選舉），主要目的是藉以評估選樣偏誤模型在不同的母體與不同的選舉情況下從事選舉預測的適用性。

貳、選舉預測方法的檢視與評估

近年來，學術界、民意調查公司、媒體、政治團體與政治人物所做的選舉預測相當多，這些選舉預測的資料來源大多是調查研究，亦即個體資料，理論建構多屬社會心理學途徑，而洪永泰（1994）、莊文忠（2000）以總體資料輔以個體資料的分析，以及盛治仁（2003）完全以總體資料作爲選舉預測的方式，是較少見的例子。同時，大多數研究皆肯定候選人因素和政黨因素是最佳的預測指標，至於其它的因素，如政見、政府施政滿意度、省籍等對選舉預測也有幫助，但是幫助相對而言較爲有限。

大多數選舉預測所使用的應變數是受訪者的投票意向，訪問題目諸如：「假如明天就是投票日，請問在（候選人姓名列舉）中，您會把票投給哪一位候選人？」或者「在（候選人姓名列舉）中，請問您最希望哪一位當選？」然而，許多受訪者在被詢問投票對象時，沒有提供確切的投票對象，尤其是距離選舉日前愈遠，則不表態的選民就愈多，因此如何來推估這些不表態者的投票意向就是選舉預測所面臨的最大挑戰。在過去的這十多年間，國內學界發展出許多選舉預測的方法，這些選舉預測方法大概可以區別爲幾類：

一、針對不表態者，以假設的投票行為法則進行推估

這類研究的特點是相信表態受訪者所回答的投票對象，而對不表態者，則假定其在選擇投票對象時有一套固定的行爲法則，只要將未表態的受訪者透過此行爲法則進行歸類，即可推估其可能的投票對象。例如劉義周（1996a）提出了三個預測模型來對未表態者加以推估：先選人再選黨，先選黨再選人，以及先詢問受訪者在政黨、候選人與政見三者之中，哪一樣最重要，然後對未回答投票意向的受訪者，用受訪者自己的法則來決定他投給哪一個候選人。第三個模型雖然在理論上較佳，但是選民不見得能清楚地說出自己的投票法則，有時候甚至會合理化自己的抉擇，例如明明是依照政黨來投票，卻回答是依照政見來投票，因此其預測效力還不如先選人再選黨的預測模型。

劉念夏（1996）提出類似的方式，只是他歸類未表態受訪者的方式，是先看未表態者是否有形象評價較高的候選人，如果有，則判定那些未表態者投給那位形象評價較高的候選人；如果沒有，則再看未表態者是否有預期某位候選人解決問題的能力高於其餘候選人，如果有，則判定其投給該候選人；如果沒有，還是無法判定，則看其政黨支持；如果未表態者有支持的政黨，則判定其投給該黨候選人，然而，對於無特定政黨支持的中立無反應者，則將其判定爲投給民進黨候選人。研究結果發現，劉念夏以 1996 年總統

選舉爲例，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，預測力可說是相當不錯。然而，何以將最後無特定政黨支持的中立無反應者判定其爲投給民進黨，並沒有提供足以說服人的理由，難以自圓其說。

盛治仁（2000）對 2000 年總統選舉提出情感溫度計作爲推估未表態受訪者的方式。所謂的情感溫度計是請受訪者對候選人給分數，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（註一）。他將未表態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高低作比較排列，總共分爲七類，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。他將最喜歡某位候選人的受訪者歸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，有不錯的預測力。

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，實則這些假定的投票模式是否可以成立，必須要經過幾次選舉的檢驗。即便是在劉義周（1996a）的研究中，我們也可以發現，同一種模型在不同的選舉中有不同的表現，很難一概而論。

二、候選人形象預測模型

梁世武（1994：115-117）認爲一般選舉預測採用直接詢問受訪者要投票給誰的方式並不十分恰當，因爲一方面許多受訪者不回答，同時有回答的受訪者也可能沒有說實話，因此他採用「候選人形象預測模型」，而所謂的候選人形象是選民對候選人特質、政見、立場所有的一種綜合態度或認知。梁世武所使用的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不錯的預測力，但是候選人形象已經綜合了我們可以想像影響選民投票行爲的所有因素，是離選民投票行爲最近的變數，自然預測效果好。選民也許不願意告訴受訪者他會投給誰，但可能會告訴受訪者他對候選人的形象評價，而選民認爲那個候選人形象比較好也幾乎就是他極可能投票的對象。這個方法也許有不錯的預測力，但卻對爲什麼有這樣的選舉結果提供有限的答案。

三、投票行為理論與統計模型的結合

將投票行爲理論與統計模型相結合，來建構影響選民投票行爲的模型，然後據以進行選舉預測是這一類研究的特色。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）。張紘炬和林顯毓（1995）以受訪者的性別、年齡、籍貫、教育程度及政黨支持作爲自變數，受訪者的投票對象爲應變數，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。延續著相同的方法，張紘炬和丁台怡（2000）及張紘炬和黃男瑋（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、籍貫爲自變數；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。

盛杏湲（1998）同樣以建立投票模型的方式來進行選舉預測，她以候選人評價、政黨認同、對政府的評價及省籍爲自變數，受訪者投票對象爲應變數，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差不多的選民，他的投票較有可能改變。該研究與其它選舉預測相當大的不同點在於：她認爲受訪者在被詢問其投票意向時可能尚未完全確定對象，有可能不回答，但也可能給予一個未經深思熟慮而當時在腦海裡閃過的候選人，將來還有變更的可能性，由於選民此一投票的不確定性及測量誤差存在的可能性，因此盛杏湲在進行選舉預測時，不以受訪者所表達的投票意向就相信他，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，則不預期投票人選。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（1999：33）認爲這樣的方式並未考慮到選民的不確定性，因爲已表態的受訪者本身已具不確定性，而未表態的受訪者，其投票行爲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爲其投票抉擇預測，等於是犧牲對於不確定性的關照。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，來作爲預測值。研究結果發現，在各縣市的預測誤差均不超過抽樣誤差，預測力相當好。

前面幾種應用統計估計模型來從事選舉預測的例子，說明選擇預測變數的重要性，無論是二元或多元對數成敗比模型、集群分析或是區辨分析，掌握適當的預測變數都是影響預測準確與否的重要關鍵。除此之外，這類研究都是以表態者所得到的估計參數來推估未表態者，然而，如果未表態者本就與表態者不同，則以表態者所得到的估計參數來推估未表態者勢必面臨預測不準的問題。

四、以總體資料為主或為輔建構模型

洪永泰以調查資料爲主，總體資料爲輔，建構預測模型，希望發揮兩者之長，互補其短，稱之爲 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（村里）的投票紀錄，再挑選指標（如各黨或候選人在過去幾次選舉的得票率）進行集群分析（cluster analysis），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）進行交叉列表，然後把每一個集群內未表態的受訪者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，換算爲有效票數的百分比，即爲其預測的得票率。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，都正確地預測到當選者，在預測得票率與實際得票率的比較上，台南市較差，差距約爲 3.1%；屏東市則相當接近，僅差距不到 0.3%。

莊文忠（2000：55-90）針對 2000 年總統選舉，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，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。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，莊文忠認爲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，會不會改爲支持其他候選人，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。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%。

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 年來台灣各項重要選舉，發現民進黨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，作爲預測 2004 年總統選舉結果的根據。研究進行的程序爲，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爲自變數，以民進黨 2000 年總統選舉地區得票率爲應變數，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爲分析單位，建構兩個模型，其中以縣市爲分析單位的模型，預測結果爲民進黨候選人可得到 49.63%的選票；以鄉鎮市爲分析單位的模型，則預測民進黨候選人得票率爲 43.03%。與 2004 年總統選舉的結果相比，以鄉鎮市爲分析單位的模型，預測力較差，以縣市爲分析單位的模型，則表現出不錯的預測力。

以總體資料來輔助推論未表態者的投票意向，基本上是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 年總統選舉爲例，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。

此外，徐永明和林昌平（2003）使用多次調查的資料，以時序模型進行選舉預測。除此之外，也有許多借用其它學科研究方法的預測模型，譬如劉文卿（1995）運用遺傳演算法的觀念，建立基因模型；李錦河和溫敏杰（1998）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；溫敏杰、杜宜軒和李錦河（2000）從統計學的角度來進行選舉預測，他們運用大數法則與不偏估計法來進行選舉預測；此外，廖達琪、景鴻鑫和楊連誠（2003）運用資訊科學的類神經網路系統軟體來從事選舉預測，也有相當的效果。

綜合以上有關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爲預測對象，其預測模型是否穩定還有待檢驗。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 模型在某些選舉的預測上並不是很準確。

一個好的選舉預測模型，不只要能夠預測得準，還必須能夠適當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，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，如此僅以表態樣本所建立的投票模型，才能夠推論到母體。問題是，受訪者的表態與否極有可能並不是隨機發生的，願意表態的人，往往有比較明顯的政治傾向，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，我們認爲若僅以這些政治態度較強烈的表態者去建立模型，容易錯估模型當中自變數對應變數的影

響力，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，如果能夠有效地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，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，應該可以得到更準確的選舉預測結果。以下我們就針對選樣偏誤問題以及其校正的方法加以說明。

參、選樣偏誤問題及其校正方法

在實驗設計中，研究者爲了要瞭解某一自變數對應變數的影響，乃隨機分派受試者到不同的組別（自變數的不同類別），因爲是隨機分派，所以可以假定不同的組別唯一的差別在於所操控的自變數，其它條件都在隨機分派下控制在一樣的情況，所以該自變數對應變數究竟有多少影響，就可以從比較不同組別在應變數的差別中得知。然而，在一般調查研究中，研究者無法隨機分派受試者到不同的組別，而是由受試者自己選擇組別，受試者可能因爲某些因素而決定他們選取的組別，如此自變數對於應變數的影響究竟有多少，可能會受到受試者在一開始選擇組別時就受到某些因素的干擾，而致無法對於自變數對應變數的影響究竟有多少，做出適當的推論，這就是在研究設計中研究者所面臨的「選擇」（selection）問題。

在作調查研究時，研究者可能面對許多必須顧慮的選擇問題，譬如：某些背景特徵的人（諸如教育程度較低、年齡較高、女性、對政治較無興趣、資訊來源較缺乏、較無政治偏好）傾向拒訪，或對某些問卷問題不回答或不表態，這些都使得研究者所得到的資料，偏向某一些背景特徵的人，使得研究者所能據以推論的是回答問卷的子樣本（sub-sample），而他們極可能與不回答問卷的子樣本是兩群不一樣的人，那麼，當以表態者去推論不表態者，選樣偏誤無法避免。真實的母體參數值應該是合併兩個子樣本之後的參數估計值，然而，問題卻正在於此，研究者根本得不到未表態子樣本的參數估計值。在這裡「加權」的方法無法奏效，因爲加權是用表態者去推估不表態者，然而如果這兩群人自始就不同，則加權可能會造成反效果。要得到正確的母體參數估計值，必須藉助選樣偏誤模型。

Heckman（1979）無疑是處理選樣偏誤的先驅者，他所使用的方式是從建構無反應過程的模型而來，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爲何願意回答的模型，便有方法可以矯正僅以有反應子樣本所建構模型的偏差（Brehm, 1993：121）。Heckman 使用了兩階段的方法來處理選樣偏誤的問題，我們首先考慮以下兩個方程式：

$$Y_{1i} = \beta_1' X_{1i} + u_{1i} \qquad (1)$$

$$Y^*_{2i} = \beta_2' X_{2i} + u^*_{2i} \qquad (2)$$

第一個方程式是結果方程式（outcome equation），也就是研究者有興趣的研究主題，在這裡，我們先考慮應變數是連續變數的狀況，因爲應變數是連續變數，因此一般使用多元迴歸來進行參數估計。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爲 0 的常態分配，且誤差項與迴歸模型中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_{1i}）將符合這個基本假定，但是實際上能夠進入結果方程式中的樣本並不是全部的樣本，而是經過一個選擇的過程，必須符合某一些標準，也就是在上述方程式（2）選樣方程式（selection equation）中，$Y^*_{2i} \geq 0$（註二），則該樣本才能夠被選入結果方程式中進行分析。然而，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，誤差項的平均數既非爲 0，且與模型中其他自變數之間的關係也並不是獨立的，因爲這個原因，模型的參數估計值會有所偏差，相關的證明可參考 Achen 一書（1986：99, 130-137）。

爲了校正上述的問題，Heckman 採兩階段校正方式，第一階段先估計選樣方程式（selection equation），也就是方程式（2），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 模型來進行分析，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）：

$$IMR_i = \varnothing(\beta_2' X_{2i}) / \Phi(\beta_2' X_{2i})$$

在上式中，$\varnothing$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）。簡單地說，IMR_i 是樣本 i 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進行參數估計的矯正。如此一來，原本的誤差項（u_{1i}）中與其它自變數相關的部分將被移除，新的誤差項將符合迴歸模型的假定，而其它原有的自變數所得到的迴歸係數也會是正確的，IMR 的迴歸係數值代表選樣方程式與結果方程式誤差項之間的相關程度，其係數值大小便代表選樣偏誤嚴重程度的大小。

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 認爲 Heckman 在第一階段的選樣方程式中使用 probit 模型來估計，雖然所得結果較爲精確，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，因爲它的

應變數不是連續變數，但是仍然將它視爲線性模型來進行估計，對於自變數的迴歸係數解釋方式是，在其它條件不變的情況下，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，不過在沒有違反最小平方法（OLS）其它基本假定的情況下，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）（註三）。LPM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（1986）首先在第一階段利用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.99或低於.01則過錄爲這兩個臨界值，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。

上面兩個方法都可以有效地校正選樣偏誤，但是這兩種方法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二分變數時，譬如像投票模型投給或不投給某候選人，其誤差項將不是呈現常態分佈，則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的（inconsistent）（Brehm, 1993：123）。爲了解決這個問題，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（$\hat{u}_{2i}$）。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：

$$Y_{1i} = \beta_1' X_{1i} + \alpha \hat{u}_{2i} (\beta_1' X_{1i}) (1 - \beta_1' X_{1i}) + v_{1i}$$

其中$\beta_1' X_{1i}$就是原本的結果方程式等號的右邊，$\hat{u}_{2i}$是第一個階段所求出的誤差項，v_{1i}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程度。

另外一個方法則是Dubin與Rivers（1989）所發展出來的方法。與Heckman和Achen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。同樣的，我們先考慮以下兩個方程式：

結果方程式爲：

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

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

選樣方程式爲：

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

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

由於Y_{1i}發生的機率是在Y_{2i}發生情況下的條件機率，我們必須考慮聯合常態分配函數（joint normal distribution funciton）。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$H(u_2)= F(\infty, u_2 ; \rho)$。函數中所列出來的$\rho$，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$G(\bullet, \bullet ; \rho)$代表$F(\bullet, \bullet ; \rho)$的右尾（upper tail）機率，例如：

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

接著計算樣本被選入的機率：

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

三種可能結果的機率如下所示：

（一）被選入的樣本，且Y_{1i}等於1

$$P(\beta_1, \beta_2, \rho)=Pr(Y_{1i}=1, Y_{2i}=1 \mid X_{1i}, X_{2i})$$
$$= Pr(Y^*_{1i}>0, Y^*_{2i}>0 \mid X_{1i}, X_{2i})$$
$$=G(-\beta_1' X_{1i}, -\beta_2' X_{2i})$$

（二）被選入的樣本，且Y_{1i}等於0

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

（三）未被選入的樣本

$$1- Q_i(\beta_2)$$

將三種可能的結果合併之後，可以得到概似函數，再將之取對數，得到如下對數概似函數（log likelihood function）：

$$LL(\beta_1, \beta_2, \rho)=\sum_{i=1}^{n} Y_{2i}Y_{1i}\ \log Pi(\beta_1, \beta_2, \rho)$$
$$+\sum_{i=1}^{n} Y_{2i}(1-Y_{1i})\log(Q_i(\beta_2)- Pi(\beta_1, \beta_2, \rho))$$
$$+\sum_{i=1}^{n}(1-Y_{2i})\log(1- Q_i(\beta_2))$$

然後用最大概似法去求得$\theta=(\beta_1, \beta_2, \rho)$的估計值。注意在此模型中，不表態的樣本也會被納入估計模型中，他們對模型估計的貢獻在於提供每一個樣本被選入或不被選入的機率。

其中ρ的估計值大小，代表了選樣偏誤的嚴重程度，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。

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。

Sheng（1994）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、社會心理學模型及 Jacobson 針對國會議員選舉的投票模型進行研究，檢驗選樣偏誤在這三種投票模型中所造成的影響。研究結果發現，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 的投票模型，選樣偏誤的情形都相當嚴重，同時該研究指出若某一自變數同時出現在選樣方程式與結果方程式，也就是說若某一影響選擇過程（投票與否）的變數同時影響到其投票對象，則該變數對投票對象的影響力特別容易被高估或低估，因此在估計投票模型時，考慮選樣偏誤確有其必要。

除了投票模型之外，選樣偏誤是一個普遍存在於許多社會科學研究中的問題，如果不加以處理，將可能造成參數估計的偏誤，甚至可能造成結果的誤判，因此，是一個相當值得關注的問題。本文以選舉預測爲研究主題，首先去估測一個校正選樣偏誤後的投票行爲模型，然後根據估計後的參數，來計算每一個樣本投票給候選人的機率，機率若超出.50，則判斷該樣本投給該名候選人。本研究使用 Dubin 與 Rivers 的二變量選樣偏誤模型，在 STATA 統計軟體中，使用指令「Heckprob」。

肆、影響選民表態與否的因素

在處理選樣偏誤的問題時，首先必須去找出影響受訪者表態與否的因素。本研究認爲受訪者不表態的情形有兩種，一種是真的沒有明確的投票對象，因此不表態；而另一種則是雖然有明確的投票對象，但是不願意說或不敢說。第一種受訪者較傾向沒有特殊的政治偏好，覺得政黨、候選人或政見對他而言都差不多，這些人一般對政治較缺乏興趣，通常也比較少會去接觸政治資訊。第二種受訪者雖然有政治偏好，但是卻對政治或選舉環境較爲敏感，對於投票對象這個相當政治敏感的問題有所保留。這種人可能是本來就傾向不回答投票對象，但是也有人可能是衡量當時的選舉環境對自己偏好的政黨或候選人不太有利，因此不願意回答投票對象。根據以上的說法推演，影響受訪者是否會表態的因素包括：是否具有政黨認同、對候選人的認知與評估情況、政治功效意識、媒體使用的情形、性別、年齡、以及教育程度。

從選民政治偏好的角度來看，如果選民沒有固定的偏好，則愈無法在候選人當中做出選擇，也使得在接受訪問時，無法給定一個明確的答案。一般來說，選民如果對某一政黨有特殊偏好，自然比較偏向投票給該政黨；反之，如果他對任何政黨都沒有特殊的偏好，則他可能會在不同的政黨與候選人之間猶豫擺盪，那麼他的投票不確定性就高，而這群選民在接受訪問時，由於還不確定自己要投給誰，因此比較傾向不表態。

候選人評價問題的回答情形對於是否表態的影響，也是影響受訪者是否有明確的投票對象。同樣從選民投票不確定性的角度出發，當選民對某一候選人有特殊的偏好時，會傾向投票給該候選人，其投票不確定性比較低，如果選民較瞭解候選人或對候選人有特殊的偏好，那麼當研究者詢問他對該候選人的評價時，選民應該會給予一個較爲明確的答案。換句話說，當選民對候選人評價沒有具體答案時，其投票不確定性較高，還無法決定要投票給誰，因此在受訪時傾向不表態。此外，候選人評價問題在首長選舉中，算是一個相當重要的問題，如果受訪者對於這個問題沒有具體的答案，表示他可能對於候選人還沒有深入的思考、比較，因此也可能導致他在訪問當下無法回答出明確的投票對象。

選民的媒體使用情形也會影響其是否表態，一方面選民如果獲取資訊愈多，就可能讓他愈確定投票的對象，因爲蒐集選舉的訊息本身就有利於選民對候選人、選舉議題、以及一般政治的瞭解，也愈有利於選民從事政治判斷與作政治抉擇。同時，資訊愈充分的人，他對政治的瞭解愈多，也會有助於去除其對政治不必要的恐懼與敏感，因此我們推論媒體使用愈多的選民，愈可能表態。

政治功效意識也是可能影響選民是否願意表態的因素。所謂的政治功效意識，根據Campbell、Gurin 和 Miller（1954：187）所提出的定義是「個人認爲其政治活動對於政治過程是有所影響，或是能夠產生影響的感覺」。也就是說，政治功效意識代表一般民眾對於自身瞭解政治的能力之評估，以及自認爲對於政策決定過程影響程度的主觀認知。Almond 與 Verba（1963）的跨國比較研究發現，如果選民自認爲自己在政治上是有能力的，那麼他在政治上的態度與行爲也會表現得愈積極。從這個觀點出發，政治功效意識愈高的選民，愈會積極表現其政治態度或行爲，在接受政治相關問題的訪問時，應該會相當地感興趣，也會比較願意表達自己在政治上的看法與意見，對於投票意向的敏感問題自然也比較不會迴避。因此，我們認爲政治功效意識愈高的選民，愈可能表態。

除此之外，性別、年齡、教育程度這三個人口變數也會對受訪者的表態與否具有影響力，一方面是影響選民的政治興趣與政治成熟度，而另一方面則是反映在對政治與選舉環境的敏感度上。在性別的影響上，女性通常較男性來得保守，且對政治較爲敏感，因此較可能不願輕易透露投票對象；而年齡愈大的受訪者，可能會因爲曾經經歷威權體制的時間較長，對民調或政治較有戒心，也比較不願意表達投票意向；至於教育程度方面的影響，教育程度的高低與政治資訊的多寡呈正相關（劉義周，1985：69），教育程度愈高的選民，愈可能有確切的投票抉擇；同時，教育程度愈低的受訪者，愈可能覺得自己懂得不多，如果貿然說出一些意見可能會被取笑，而導致較不願意表達自己的投票意向。從過去的研究結果來看，女性的未表態比例要比男性來得高一些；年齡愈高的選民愈不願意透露其投票意向；教育程度愈低的民眾愈不願意表態（盛治仁，2000：79-80；鄭夙芬和陳陸輝，2000）。

由於資料蒐集上的限制，因此在每一次選舉預測模型中，並不一定包含上列所有的變數，而是以該筆資料中既有的變數爲主，所有的資料來源與變數測量方式見於附錄。在此特別需要說明的是我們在作模型估計時，將所有的變數皆轉化成爲 0 到 1 的尺度，如此一來，所有的自變數每一單位的變動都是從最小值到最大值的變動，也就是全距的變動，以方便比較各自變數對於表態與否的影響力大小。同時，由於跨不同選舉年調查資料所使用的測量方式不完全相同（如候選人評價回答的情形變數，有的選舉全距爲 1（如台北市長選舉），有的選舉全距爲 4（如 2004 年總統選舉），因此將所有變數化爲 0 到 1 的尺度，也有助於比較跨不同年度各變數的影響力是否穩定。表 1 所示爲歷次選舉中影響選民表態與否的probit分析。應變數爲表態與否，表態設爲 1，不表態設爲 0。自變數中，是否具有政黨認同、性別、教育程度是類別變數，因此以虛擬變數的方式放入模型中，另外候選人評價問題回答情形、政治功效意識、媒體使用情形、年齡是以連續變數的方式放入模型中。

從表 1 得知，影響受訪者是否表態的最重要因素是候選人認知、評價與政黨認同，那些愈能夠區別候選人差別的，以及有政黨認同的受訪者，愈可能表態；政黨認同是中立或無反應的，比那些具有政黨認同的，有相當明顯的不表態傾向。相當值得注意的是泛藍認同者與泛綠認同者之間的差距，泛藍與泛綠認同者，雖不見得在每一次選舉都顯現出有差異，但差異的方向值得注意。在兩次總統選舉中，泛藍認同者不見得比泛綠認同者傾向表態（未達到顯著差異水準），在 2001 年北縣與 2002 年高市選舉，泛藍選民比泛綠選民傾向不表態（達到統計上的顯著差異水準），這似乎與一般的印象不完全符合，因爲一般認爲泛藍支持者比泛綠支持者傾向於表態，但注意此一估計值是在模型中已經控制了教育程度、性別、年齡、媒體使用、政治功效意識等反映個人政治成熟度與敏感度等因素之後。也就是說在控制了政治成熟度與敏感度等因素之後，泛藍或泛綠認同者是否表態與當時政黨或候選人的聲勢有關，如果當時某一黨的聲勢較高，則該黨的支持者較有表態的傾向，反之另一黨的支持者表態的傾向較弱。我們可以發現在 2001 年北縣與 2002 年高市選舉泛藍選民比泛綠選民傾向不表態，在 2002 年北市選舉泛綠認同者比泛藍認同者傾向不表態（在統計顯著水準邊緣）。由此我們可以推論受訪者會感受選舉當時的民意氣候，當發現自己偏好的政黨或候選人居於劣勢時，比較傾向保持沈默。至於在 2000 年與 2004 年總統選舉，候選人究竟誰居於優勢或劣勢較不明顯，因此泛藍與泛綠受訪者是否表態並沒有顯出統計上有意義的差距。

表 1　歷次選舉選樣方程式 probit 模型估計

	2000 總統	2001 北縣	2002 北市	2002 高市	2004 總統
政黨認同（泛綠＝0）					
泛藍	-.08(.11)	-.44(.11)***	.40(.26)	-.30(.18)*	-.22(.14)
中立無反應	-.80(.11)***	-1.24(.11)***	-1.24(.20)***	-1.25(.18)***	-1.51(.13)***
候選人認知與評價	1.16(.13)***	1.02(.16)***	.73(.17)***	1.21(.15)***	1.25(.21)***
媒體使用	.67(.16)***	.28(.13)*	.04(.24)	.48(.19)*	.28(.19)
政治功效意識			.29(.22)	.36(.16)*	
性別（女性＝0）					
男性	.31(.08)***	.25(.08)*	-.13(.17)	.14(.13)	.24(.10)*
年齡	-1.21(.24)***	.06(.25)	-.00(.50)	-.60(.39)	-.21(.32)
教育程度（低教育程度＝0）					
中教育程度	.07(.11)***	.40(.11)***	-.12(.28)	.27(.17)	-.01(.15)
高教育程度	.25(.11)*	.54(.11)***	.06(.27)	.20(.18)	.39(.15)*
常數	-.54(.23)	-.17(.22)	1.13(.42)**	-.02(.28)	-.34(.31)
分析個數	1389	1608	773	758	934
正確預測率（%）	78.40	82.40	92.37	87.34	80.51
-2 Log Likelihood	1319.57	1278.54	299.95	482.59	799.45

說明：1.應變數若表態則爲 1；不表態則爲 0。

2.表中數字爲 probit 模型分析的估計值，括弧中的數字爲標準誤。

3.***: $p < .001$；**: $p < .01$；*: $p < .05$；$: $p < .1$。

資料來源：見於附錄。

除此之外，媒體使用、政治功效意識、性別、年齡、教育程度等變數跨不同的選舉，對於民眾表態與否的影響，並不完全一致，比較值得注意的是，對於 2002 年台北市長選舉而言，這些因素都不影響選民是否表態。主要的原因可能是因爲一個超人氣的候選人馬英九對上一個相對上較弱的候選人李應元，以致於相當多數民眾都願意說他要將票投給馬英九，即使是那些一般對政治較敏感或政治上較不那麼成熟的民眾，似乎表態說要投給馬英九對他們不會造成任何的負擔，以致於不表態民眾被壓縮到只剩下大約 10 個百分點左右（在這波民調中，有 64.63%的民眾說要投給馬英九，而 24.14%的民眾說要投給李應元，不表態的僅有 11.23%的民眾）。如此使得應變數是一個極偏的分配，以致於估計可能比較不穩定。

若排除掉2002年的台北市長選舉之後，其它選舉年度的發現就相當一致了：使用媒體愈多的愈傾向表態，政治功效意識愈高者愈傾向表態、男性比女性傾向表態，教育程度高者較傾向表態。年齡的影響雖然大多未達統計上的顯著水準，但大致而言，年齡愈長愈傾向不表態。

伍、投票模型的估計與選舉預測

在投票對象的結果方程式方面，本研究將採用一般投票行爲研究中眾所認爲最主要的變數來建構模型，包括候選人評價、政黨認同、對現任者施政滿意度、省籍因素，除此之外，在總統選舉中多加入統獨立場變數。

「候選人評價」這個因素在投票行爲上一向扮演相當重要的角色（Rahn et al., 1993），有學者認爲近年來在美國的選舉中，競選的方式已經逐漸走向以「候選人爲中心」（candidate-centered），而不是以往「以政黨爲中心」（Asher, 1988）。而在國內許多的研究中也肯定候選人因素在影響選民投票抉擇上扮演相當重要的角色（梁世武，1996；黃秀端，1997；盛杏湲，1998；盛治仁，2000）。在研究範圍中的這五次選舉，候選人所獲得的注意程度都非常高，只要一有候選人的任何消息，不論是電子媒體或是平面媒體都會競相報導，如此高的曝光率，使選民會更加注意候選人本身，連帶使得候選人評價這個因素在選舉中必定也會扮演相當重要的角色，那些愈偏向給某位候選人較高評價的選民，愈傾向把票投給該位候選人。

政黨認同一向被認爲是影響選民投票行爲的重要變數。所謂政黨認同指的是心理層面的認同，個人依附於某一個政黨，對政黨的一種歸屬感與忠誠感（Campbell et al., 1960：121），Campbell 等人也發現美國選民的政黨認同具有很高的穩定性（Campbell et al., 1960：150），同時在他們所提出的「漏斗狀因果模型」（funnel of causality）的分析架構，政黨認同被視爲是貫穿影響選民投票行爲的主軸。在國內學者的相關研究中，多肯定政黨認同對於選民的投票抉擇有相當的影響力（徐火炎，1993；何思因，1994；陳義彥，1994；劉義周，1996b；盛杏湲，1998、2002；陳陸輝，2002）。

放入現任者施政滿意度這個變數，是根據「回溯性投票」（retrospecitve voting）的論點而來的，Fiorina（1981）認爲一個理性的選民是評估過去執政黨的表現來決定他的投票對象，而非依據預期未來的表現，如果現任的政黨表現好，則會繼續投票支持執政黨，如果現任的政黨表現不佳，則選民會將選票投給在野黨。

除了上述幾個因素之外，省籍因素亦將納入模型之中。從過去威權時代至民主化以後，台灣各族群之間的權力消長歷史經驗來看，可分爲「本省族群」及「外省族群」，

也就是一般所謂的「本省人」與「外省人」（劉子立，2004：5）。在威權時期，由於政治上長期是外省人所壟斷的局面，本省人屬於弱勢的族群，因此造成本省人產生排除外省族群的意識，反應到投票抉擇上，本省人便比較會投票給本省籍的候選人或政黨，而民主化之後，外省族群長期以來在政治上的優勢地位被打破，促使他們產生危機意識，反應到投票抉擇上，便會投票給外省籍的候選人或政黨。陳義彥在研究 1992 年的立委選舉時，發現省籍是區別選民不同集群的最重要向度（陳義彥，1994：31）。廖益興針對 1996 年總統大選所進行的研究也發現，省籍是影響選民支持李登輝與否的重要因素（廖益興，1996：198-199）。盛杏湲的研究也同樣指出，省籍從過去到現在都是台灣重要的政治分歧，對選民投票抉擇的影響相當重要（盛杏湲，1998：49；盛杏湲，2002：56）。

由於長久以來兩岸分治的事實，再加上每到選舉，統獨議題會被政黨或政治人物提出，民眾自然而然不需要花費太大的成本，便可以很輕易地辨識自己及主要政黨的統獨立場（盛杏湲，2002）。至於選民的統獨立場對於其投票抉擇的影響，在台灣的研究中呈現出並不一致的結果。游盈隆（1994）針對第二屆國大選舉所做的研究當中發現，統獨問題對於選民投票幾乎沒有影響力；在陳義彥（1994）對第二屆立委選舉所做的研究中，他以政黨形象、政黨表現、候選人形象、統獨政見、族群意識、省籍等變項對選民作集群分析，並討論各集群與實際投票行爲之間的關聯性，結果發現統獨立場並不是區別選民不同集群的重要向度；劉義周（1996b：16）的研究指出，在 1996 年的總統選舉中，選民的統獨立場對於他是否投票給李登輝並沒有明顯的影響；但是謝復生等人（1995）所做的研究則指出，不同議題會左右台灣選民對政黨的評價，而選民對政黨的評價又會影響其投票抉擇，而這些議題當中就包含了統獨議題；盛杏湲（2002）的研究亦顯示，在九〇年代的五次選舉當中，統獨議題對選民的投票對象而言都是具有顯著影響的變數，只是在立委選舉中統獨立場對於選民的影響較小。從以上學者們的研究來看，我們可以發現選民的統獨立場對於其投票抉擇的影響並不穩定，但是從理論的角度來看，它的確也是可能影響選民投票抉擇的因素之一，因此仍然必須將之放入模型當中一起討論。不過因爲統獨議題是屬於全國性的議題，本研究預期此因素僅會在全國性選舉中發揮效用，因此在地方性選舉的模型中，並不放入此因素。

爲了比較有無處理選樣偏誤，對於參數估計所造成的影響，將建立兩個模型，一是不處理選樣偏誤，也就是僅用已表態者的投票對象建立模型，將它稱爲「傳統 probit 模型」，二是校正選樣偏誤後建立的模型，稱爲「選樣偏誤模型」。模型估計的結果見於表 2 至表 6。我們舉表 2 所顯示的 2000 年總統選舉爲例加以說明。

表 2　2000 年總統選舉選民投票模型估測

	選樣偏誤模型（校正選樣偏誤）	傳統 probit 模型（未校正選樣偏誤）
結果方程式（投陳水扁＝ 1，投非陳水扁＝ 0）		
候選人評價(陳水扁最高＝ 0)		
連戰最高	-2.074(.306)***	-2.344(.313)***
宋楚瑜最高	-2.797(.334)***	-3.138(.316)***
宋扁同，連戰最低	-1.266(.207)***	-1.357(.223)***
連扁同，宋楚瑜最低	-.994(.205)***	-1.090(.220)***
連宋同，陳水扁最低	-2.737(.486)***	-3.119(.536)***
無法比較高低	-1.286(.199)***	-1.594(.186)***
藍綠認同(泛綠＝ 0)		
泛藍	-1.650(.183)***	-1.857(.165)***
中立無反應	-.440(.174)**	-.823(.153)***
省籍(外省＝ 0)		
本省	.240(.201)	.310(.237)
統獨立場(傾向獨立＝ 0)		
傾向統一	-.620(.198)**	-.729(.220)**
維持現狀	-.360(.154)**	-.415(.173)*
常數	2.068(.277)***	2.059(.313)***
ρ	-.731(.161)***	
選樣方程式(表態＝ 1，不表態＝ 0)		
性別（女性＝ 0）		
男性	.273(.081)**	
年齡	-1.295(.236)***	
教育程度（低教育程度＝ 0）		
中教育程度	.063(.104)	
高教育程度	.264(.109)*	
媒體使用情形	.591(.162)***	
政黨認同（泛綠＝ 0）		
泛藍	-.082(.107)	
中立無反應	-.822(.105)***	
候選人評價回答情形	1.128(.129)***	
常數	-.395(.235)$^{\$}$	
分析個數	1389	963
-2 Log Likelihood	1800.2396	486.5767
LR test（ρ＝ 0）	5.90*	

說明：1.表中數字爲 probit 模型分析的估計值，括弧中的數字爲標準誤。

2.***: $p < .001$；**: $p < .01$；*: $p < .05$；$^{\$}$: $p < .1$。

表 3　2001 年台北縣長選舉選民投票模型估測

	選樣偏誤模型 (校正選樣偏誤)	傳統 probit 模型 (未校正選樣偏誤)
結果方程式(投蘇貞昌＝ 1，投王建煊＝ 0)		
候選人評價(蘇貞昌較高＝ 0)		
王建煊較高	-2.473(.198)***	-2.582(.175)***
無法比較高低	-1.294(.148)***	-1.308(.149)***
蘇貞昌施政滿意度	1.939(.383)***	2.116(.374)***
政黨認同(泛綠＝ 0)		
泛藍	-1.656(.159)***	-1.657(.163)***
中立無反應	-.935(.216)***	-.665(.184)***
省籍(外省＝ 0)		
本省	.516(.168)**	.555(.173)***
常數	.463(.338)	.458(.355)
ρ	.480(.226)***	
選樣方程式(表態＝ 1，不表態＝ 0)		
性別(女性＝ 0)		
男性	.264(.079)**	
教育程度(低教育程度＝ 0)		
中教育程度	.358(.099)***	
高教育程度	.503(.101)***	
媒體使用情形	.287(.131)*	
政黨認同(泛綠＝ 0)		
泛藍	-.425(.104)***	
中立無反應	-1.235(.106)***	
候選人評價回答情形	1.021(.153)***	
常數	-.157(.184)	
分析個數	1630	1285
-2 Log Likelihood	1830.836	520.1994
LR test(ρ＝ 0)	3.13 $^{\$}$	

說明：1.表中數字爲 probit 模型分析的估計值，括弧中的數字爲標準誤。

2.***: $p < .001$；**: $p < .01$；*: $p < .05$；$^{\$}$: $p < .1$。

表 4　2002 年台北市長選舉選民投票模型估測

	選樣偏誤模型 (校正選樣偏誤)	傳統 probit 模型 (未校正選樣偏誤)
結果方程式(投李應元＝ 1，投馬英九＝ 0)		
候選人評價(李應元較高＝ 0)		
馬英九較高	-1.600(.376)***	-1.751(.399)***
無法比較高低	-.292(.232)	-.426(.242)$
馬英九施政滿意度	-3.479(.560)***	-3.922(.565)***
藍綠認同(泛綠＝ 0)		
泛藍	-2.085(.264)***	-2.147(.280)***
中立無反應	-.927(.341)**	-1.567(.257)***
省籍(外省＝ 0)		
本省	.429(.282)	.499(.311)
常數	2.690(.407)***	2.915(.430)***
ρ	-.841(.186)	
選樣方程式(表態＝ 1，不表態＝ 0)		
政黨認同(泛綠＝ 0)		
泛藍	.432(.235)$	
中立無反應	-1.190(.194)***	
候選人評價回答情形(無＝ 0)		
有回答	.764(.154)***	
常數	1.194(.183)***	
分析個數	789	717
-2 Log Likelihood	500.8228	178.5579
LR test(ρ＝ 0)	2.37	

說明：1.表中數字爲 probit 模型分析的估計值，括弧中的數字爲標準誤。

2.***: $p < .001$；**: $p < .01$；*: $p < .05$；$: $p < .1$。

表 5　2002 年高雄市長選舉選民投票模型估測

	選樣偏誤模型 (校正選樣偏誤)	傳統 probit 模型 (未校正選樣偏誤)
結果方程式(投謝長廷＝ 1，投黃俊英＝ 0)		
候選人評價(謝長廷較高＝ 0)		
黃俊英較高	-2.427(.293)***	-2.643(.290)***
無法比較高低	-1.060(.268)***	-1.342(.282)***
謝長廷施政滿意度	1.897(.467)***	2.132(.502)***
藍綠認同(泛綠＝ 0)		
泛藍	-1.426(.261)***	-1.610(.264)***
中立無反應	.141(.268)	-.296(.278)
省籍(外省＝ 0)		
本省	.344(.317)	.352(.357)
常數	.629(.439)	.530(.489)
ρ	-.819(.167)***	
選樣方程式(表態＝ 1，不表態＝ 0)		
媒體使用情形	.431(.182)*	
政黨認同(泛綠＝ 0)		
泛藍	-.252(.175)	
中立無反應	-1.264(.170)***	
候選人評價回答情形(無＝ 0)		
有回答	1.233(.146)***	
政治功效意識	.574(.156)***	
常數	-.086(.222)	
分析個數	771	621
-2 Log Likelihood	655.2378	159.4415
LR test(ρ＝ 0)	3.47 ^{s}	

說明：1.表中數字爲 probit 模型分析的估計值，括弧中的數字爲標準誤。

2.***: $p < .001$；**: $p < .01$；*: $p < .05$；^{s}: $p < .1$。

表 6　2004 年總統選舉選民投票模型估測

	選樣偏誤模型 (校正選樣偏誤)	傳統 probit 模型 (未校正選樣偏誤)
結果方程式(投陳呂＝ 1，投連宋＝ 0)		
候選人評價(陳呂較高＝ 0)		
連宋較高	-1.707(.317)***	-2.154(.337)***
無法比較高低	-.684(.275)*	-1.151(.323)***
陳水扁施政滿意度	1.769(.514)**	2.351(.630)***
藍綠認同(泛綠＝ 0)		
泛藍	-1.864(.317)***	-2.281(.359)***
中立無反應	-.664(.313)*	-1.584(.316)***
省籍(外省＝ 0)		
本省	.922(.377)*	1.000(.488)*
統獨立場(傾向獨立＝ 0)		
傾向統一	-.326(.348)	-.479(.486)
維持現狀	-.684(.254)**	-.834(.321)**
常數	.980(.472)*	.890(.620)
ρ	-.938(.125)***	
選樣方程式(表態＝ 1，不表態＝ 0)		
性別(女性＝ 0)		
男性	.209(.096)*	
教育程度(低教育程度＝ 0)		
中教育程度	.036(.127)	
高教育程度	.494(.131)***	
媒體使用情形	.307(.177)$^{\$}$	
政黨認同(泛綠＝ 0)		
泛藍	-.235(.136)$^{\$}$	
中立無反應	-1.543(.128)***	
候選人評價回答情形	1.121(.194)***	
常數	-.335(.240)	
分析個數	951	647
-2 Log Likelihood	921.8480	114.7517
LR test(ρ＝ 0)	4.83*	

說明：1.表中數字爲 probit 模型分析的估計值，括弧中的數字爲標準誤。

2.***: p <.001；**: p <.01；*: p <.05；$^{\$}$: p <.1。

在2000年總統選舉的選樣偏誤模型中，結果方程式的應變數是選民投票對象，以投給陳水扁爲1，非投給陳水扁爲0，之所以僅簡單地作二分法，是因爲研究者現有的統計軟體只能支應結果方程式爲二分變項，而無法處理多分變項，同時，在理論上，要區分選民投票給連戰或宋楚瑜，涉及到棄保效應，在未能充分掌握棄保效應的測量變數時，暫且不將投給連宋的選民分開。自變數除了候選人評價、政黨認同及省籍之外，還加上統獨立場，其中候選人評價、政黨認同、省籍、統獨立場都是類別變數，我們將以虛擬變數的方式放入模型中，候選人評價區分爲對連戰評價最高、對宋楚瑜評價最高、對陳水扁評價最高、宋扁同，對連戰評價最低、連扁同，對宋楚瑜評價最低、連宋同，對陳水扁評價最低、以及無法比較高低等七類，比較基礎爲對陳水扁評價最高；政黨認同有認同泛藍、認同泛綠及中立無反應三類，比較基礎爲認同泛綠；省籍區分爲本省人與外省人，比較基礎爲外省人；統獨立場有傾向獨立、傾向統一、維持現狀三類，比較基礎爲傾向獨立。同樣的，與前述選樣方程式的估計相同，我們也將變數都化爲0到1的尺度，以方便比較各自變數的影響力大小。

首先我們可以發現到在表2中，各變數的係數方向都與理論預期一致，而且除了省籍外，每個變數的估計值都達到統計上的顯著水準。ρ等於-.731，ρ小於0的意思是在不考慮選樣偏誤的情況下，將低估結果方程式中Y=1發生的機率，也就是說將低估選民投票給陳水扁的機率。同時值得注意的是，選樣偏誤模型的分析樣本數高於傳統 probit 模型，主要原因是選樣偏誤模型囊括了表態與不表態兩個子樣本，而對於不表態的樣本，即使他在結果方程式的某一變數爲缺失值，他仍然可能進入模型中。而在傳統估計模型中，唯有表態且在結果方程式中所有變數都不缺失的樣本才能進入分析中。

觀察兩個模型中參數估計值的差異狀況，在候選人評價方面，可以發現傳統 probit 模型高估了候選人評價的影響，幾乎在候選人評價的每一個類別，相對於評價陳水扁最高的類別，傳統 probit 模型中的係數都比選樣偏誤模型的係數來得高，譬如給連戰最高評價者相對於給陳水扁最高評價者，校正選樣偏誤後的參數估計值從-2.344 減弱爲-2.074，意味著在未校正選樣偏誤的情況下，模型高估給連戰最高評價者與給陳水扁最高評價者之間的差距；傳統 probit 模型也高估了給宋楚瑜最高評價者與給陳水扁最高評價者之間的差距，而且高估的程度蠻多的，表示在校正選樣偏誤之後，兩者的差距其實沒有那麼大。同時，在無法比較候選人評價高低者的估計值顯示，未校正選樣偏誤的傳統 probit 模型高估了無法比較候選人評價高低者與給陳水扁最高評價者之間的差距，係數由-1.594 減少爲-1.286，表示在校正選樣偏誤之後，無法比較候選人評價者與給陳水扁最高評價者的距離減少了.308，變動幅度還算蠻大的，由於這個變數是虛擬變數，因此我們可以利用表中變數的參數估計值，分別計算出無法比較候選人評價高低者與給連

戰最高評價者、給宋楚瑜最高評價者之間的距離，若以給連戰最高評價者爲比較基礎的話，在傳統probit模型中，無法比較候選人評價高低者的參數估計值是.75（-1.594+2.344），在選樣偏誤模型中的參數估計值是.788（-1.286+2.074），表示在校正選樣偏誤之後，給連戰最高評價者與無法比較候選人評價高低者之間的差距增加了一點點，若以給宋楚瑜最高評價者爲比較基礎的話，在傳統 probit 模型中，無法比較候選人評價高低者的參數估計值是 1.544（-1.594+3.318），在選樣偏誤模型中的參數估計值是 1.511（-1.286+2.797），表示在校正選樣偏誤之後，給宋楚瑜最高評價者與無法比較候選人評價高低者之間的差距減少了一點點。綜合來看，在校正選樣偏誤之後，給予連戰或宋楚瑜最高評價者與無法比較候選人評價高低者的距離沒有什麼變動，而無法比較候選人評價者與給陳水扁最高評價者的距離減少，表示在無法比較候選人評價高低的選民中，不表態選民的投票傾向比表態選民更接近給陳水扁評價較高者，也就是說會比較傾向投票給陳呂，因此若不校正選樣偏誤的話，將低估陳呂的得票率。

在政黨認同參數估計值的變化上，在沒有校正選樣偏誤的情況下，傳統 probit 模型高估了認同泛藍的選民比認同泛綠的選民更傾向不投票給陳呂的強度，未校正選樣偏誤時的參數估計值爲-1.857，校正選樣偏誤之後的參數估計值則減少爲-1.650，產生這樣的變化，與本研究的預期是相符的，因爲我們僅以表態者建立傳統 probit 模型，而有政黨認同的選民又比沒有政黨認同的選民更傾向表態，因此使得傳統 probit 模型高估藍綠認同者之間的差異；接下來我們觀察到，未校正選樣偏誤的傳統 probit 模型高估了中立無反應選民比認同泛綠選民更傾向不投票給陳呂的強度，係數由-.823 減少爲-.440，表示在校正選樣偏誤之後，中立無反應者與泛綠的距離減少了.383，是個不小的變動，由於這個變數是虛擬變數，因此我們可以利用表中變數的參數估計值計算出中立無反應者與認同泛藍者的距離，若以認同泛藍爲比較基礎的話，在傳統 probit 模型中，中立無反應者的參數估計值是 1.034（-.823+1.857），在選樣偏誤模型中的參數估計值是 1.21（-.440+1.650），表示在校正選樣偏誤之後，泛藍與中立無反應之間的差距變大了。綜合來看，校正選樣偏誤之後，泛藍與中立無反應的距離變大，泛綠與中立無反應的距離變小，表示在政黨認同是中立無反應的選民中，不表態者與表態者相比，其投票傾向與認同泛綠者比較接近，會比較傾向投票給陳呂，因此若不校正選樣偏誤的話，將低估陳呂的得票率。

在統獨立場方面，參數估計值的變動並不算太大，傳統 probit 模型高估傾向統一比傾向獨立的選民更傾向不投票給陳呂的強度，校正選樣偏誤後的參數估計值由-.729 減小爲-.620，而在校正選樣偏誤之後，傾向維持現狀的選民比傾向獨立的選民更傾向不投票給陳呂，其強度也是減弱的。省籍參數估計值的變動相當小，在校正選樣偏誤之後，省

籍的參數估計值從.310 減弱到.240，僅減少了.07。

從以上這幾個變數的估計值變動程度來看，在 2000 年的總統選舉中，校正選樣偏誤後，政黨認同是中立無反應者的變動程度是最多的，而政黨認同是中立無反應的選民也正是在訪問中非常不願意表態的一群人，這表示如果我們將不願意表態的人忽略不計，便容易錯估政黨認同是中立無反應者的真正投票傾向，當然就會進一步去影響到候選人得票率的預測。無法比較候選人評價高低變數校正選樣偏誤之後的變動程度也不小，由於對候選人評價問題的回答情形影響了是否表態，對候選人評價問題都回答的選民較傾向表態，在都回答的情況下，才有可能分出對候選人的評價高低，因此僅以表態者所建立的傳統 probit 模型，便可能高估給不同候選人較高評價者之間的差異。而省籍影響力的變動最少，可能是因爲省籍本身並不太會去影響是否表態，因此它的參數估計值沒有太大的變化。

根據以上兩個模型的參數估計所計算出的預測結果，詳見表 7。從表 7 中可以得知，傳統 probit 模型的預測結果，在沒有考慮選樣偏誤的問題之下，果然低估了陳呂的得票率，與實際得票率的誤差達到 7.68%，表現得甚至比直接電話訪問的結果更差，反觀選樣偏誤模型的表現則相當不錯，與實際得票率的誤差只有.59%，獲得比傳統 probit 模型更好的預測結果。

綜觀五個選舉的模型估計與預測結果，有幾點值得注意的發現：首先，估計模型跨不同的選舉呈現相當的穩定度，每一個自變數對於投票對象的影響方向，皆如理論所預期的，亦即，愈是對於某候選人評價愈高，愈可能投給該候選人，愈是對現任者滿意，則愈傾向投給現任者或現任者的政黨候選人，而愈是認同某一陣營的政黨，則愈傾向投給該陣營的候選人。此外，本省籍比外省籍傾向投給民進黨的候選人，而統獨立場愈是傾向獨立的，則愈可能投票給民進黨的候選人。

而更加值得注意的是，當比較校正選樣偏誤與未校正的傳統模型時，從五個選舉的模型中都可以相當清楚地看見，當我們未校正選樣偏誤時，可能高估給予不同候選人最高評價者之間的差距、也高估藍綠政黨認同者之間的差距、以及高估統獨支持者之間的差距，也就是說會高估這些變數對於投票抉擇的影響；同時也會高估了政府首長施政滿意度與省籍對於投票抉擇的影響。此一發現相當一致與穩定，顯現出因爲願意回答自己投票對象的受訪者，往往是政治偏好相當確定且較強烈的選民，當我們在估計模型時，若只掌握了這些政治偏好較強或較確定的表態者，而忽略政治偏好較弱或較不明確的不表態者，我們極可能高估自變數對應變數的影響，尤其是一般認爲影響選舉結果最重要的候選人因素與政黨因素。

表 7　歷次選舉電訪表態、傳統 probit 模型、選樣偏誤模型與實際得票率的比較

	實際得票率	電訪表態者	傳統 probit 模型	選樣偏誤模型
2000 年總統選舉[a]				
投陳呂	39.61	35.68	31.93	40.20
不投陳呂	60.39	64.32	68.07	59.80
與實際誤差	——	3.93	7.68	0.59
2001 年台北縣長選舉				
蘇貞昌	51.58	53.01	56.24	52.29
王建煊	48.42	46.99	43.76	47.71
與實際誤差	——	1.43	4.66	0.71
2002 年台北市長選舉				
李應元	35.89	27.19	27.43	27.43
馬英九	64.11	72.81	72.57	72.57
與實際誤差	——	8.70	8.46	8.46
2002 年高雄市長選舉				
謝長廷	51.66	43.39	49.72	52.82
黃俊英	48.34	56.61	50.28	47.18
與實際誤差	——	8.27	1.94	1.16
2004 年總統選舉				
扁呂	50.11	46.81	43.92	50.36
連宋	49.89	53.19	56.08	49.64
與實際誤差	——	3.30	6.19	0.25

說明：a. 爲使各次選舉都爲兩方對決，因此 2000 總統選舉以投陳水扁爲 1，非投陳水扁爲 0，其餘各次選舉皆以前兩名候選人爲分析對象，投給其他候選人者，以缺失資料處理。實際得票率的計算，係排除掉第三名及以外的候選人之後，兩方對決的相對比例。

同時值得注意的是，除了在台北市的選樣偏誤估計模型中，ρ不顯著之外，其餘四次選舉都顯著。ρ代表選樣方程式的誤差項與結果方程式誤差項之間的相關係數值，ρ大於 0 的意思是，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，也就是說，不表態者傾向於投票給非民進黨；反之，ρ小於 0 的意思是，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，也就是說，不表態者傾向投票給民進

黨；ρ若等於 0，表示不表態者與表態者在投票傾向上並沒有不同，至少在當下模型掌握的變數裡面，無法指出究竟不表態者與表態者有所不同。在五次選舉中，有三次選舉（2000 年總統選舉、2002 年高雄市長選舉、2004 年總統選舉）ρ小於 0，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，也就是說不表態者傾向投票給民進黨，這也是我們一般的認識，亦即民意調查的結果通常低估民進黨的得票率。但是在 2001 年台北縣長選舉估計模型ρ大於 0，表示在該次選舉裡，不表態者傾向投票給非民進黨，也許是因爲在該次選舉裡，由於民進黨候選人爲現任縣長蘇貞昌，相當具有選戰的優勢，而泛藍陣營的候選人王建煊是在選舉日前不久才確定，而在整個選舉的過程裡，聲勢較弱，因此傾向投票給王建煊的受訪者比較不願表態，以致於顯示出不願表態者傾向投票給非民進黨的候選人，若不校正選樣偏誤，則可能會低估王建煊的得票。至於在 2002 年台北市長的選舉估計模型，ρ大於 0，但是無法拒絕ρ=0 的假設，表示該年度不表態者與表態者沒有明顯不同的投票傾向。

從表 7 可以發現，除台北市之外，在其餘四次選舉，校正選樣偏誤之後，所做的選舉預測皆相當準確，其最大的誤差值不過 1.16%，皆在抽樣誤差的範圍內，且都比不校正選樣偏誤還要準確。然而，對 2002 年台北市長選舉的估計，選樣偏誤模型發揮不了校正的功能，以致於校正與否得到的預測結果完全一樣，都與實際的選舉結果有相當大的誤差。我們認爲之所以會得到這樣的結果，可能的原因有下列幾個：第一，在選樣方程式的設定上，可能沒有掌握到真正影響表態或不表態的原因，事實上在選樣方程式估計上，我們可以發現，性別、年齡、教育程度、媒體使用情形等變數，對於是否表態的影響都不顯著，整個選樣方程式到最後僅剩下三個變數，這樣的模型設定，有可能並無法掌握到真正影響表態或不表態的因素。

第二，表態者所給予的答案可能並非是真實的，在兩個候選人實力懸殊，選舉勝負早已底定的情況下，某些傾向投票給李應元的受訪者，可能因爲投票意向與主流趨勢不符，所以不願意回答投票對象，甚至欺騙他會投票給馬英九，因此造成估計模型的偏差，從而造成選舉預測的偏誤，而如何判斷哪些因素影響受訪者真實或虛假的回答，是調查研究中研究者必須去重視的問題，本研究由於缺乏適當的測量變數，暫不對此作探討。

第三，也許民意調查本身的自毀效果使然，在選前許多民意調查皆顯示馬英九以極高的比例勝李應元，致使某些原本支持馬英九的受訪者發現馬英九的選情大好，一定選勝，就不去投票；反之，傾向民進黨的選民不希望選舉太過懸殊而去投票的意願轉強，造成選舉結果並沒有如調查資料預期的馬英九大勝，李應元大敗的懸殊結果。

陸、結論

本文的主要目的在於評估選樣偏誤對於投票模型的估計所造成的影響，並且試圖藉由矯正選樣偏誤所造成的問題，得到較正確的參數估計值，並進而作更精確的選舉預測。在本文中，我們採取 Dubin 與 Rivers 所發展出來的選樣偏誤模型爲研究方法，爲了檢視選樣偏誤模型在選舉預測上的穩定性，我們將之應用在五次選舉中。結果發現在五次選舉中，未校正選樣偏誤都會造成高估自變數對應變數的影響，以致於造成選舉預測的偏誤。當我們校正選樣偏誤後，在四次選舉中都發揮了極好的效果，預測的誤差都比原本不校正選樣偏誤來得更小，且誤差都不超過 1.16%，可謂相當地準確。唯有在一次選舉無法發揮校正的效果，但是即便如此，也並不會比不校正更差。我們認爲這樣的效果顯示，選樣偏誤模型是一個相當可以信賴的選舉預測工具。

本文以選舉預測爲例來探討選樣偏誤模型的適用性，有一個意義是因爲選舉預測提供了我們一個將學術的理論與方法應用於實際政治現象的機會，由於選舉結果可以在選後揭曉，因此校正選樣偏誤的效果如何，也可以經由比較選舉預測的結果與正式的選舉結果而得知。由於本研究的結果相當一致而穩定，因此我們得知如果不校正選樣偏誤，則研究者相當可能只掌握到一部份受訪者的答案，而忽略掉另一些受訪者的答案，從而高估或低估自變數對於應變數的影響。

民意調查一向被認爲是反映民意的重要工具，有些人甚至認爲由於民意調查並非自我選擇（self-selected），因此民意調查所得到的民意，可能比一般政治參與行動所表達的民意更具有代表性（Verba, 1996）。然而，選樣偏誤問題的提出，說明了在某些民意調查的主題上，其實受訪者會自己選擇接受訪問或拒絕訪問，回答或不回答，表態或不表態，如果這個選擇並非隨機而是依據某些規則而產生的，則我們有理由相信如此得到的民意結果可能是有所偏誤的，也因此，選樣偏誤模型應該可以相當廣泛地運用在許多民意調查的主題中。譬如針對某些較難回答的民意調查問題，可能只有教育程度較高或較成熟練達者能回答，而如果他們的意見偏向某一個方向，則僅以他們的意見推論到全部民眾的意見，勢必造成偏誤。又或者針對某一項社會有可欲標的的議題，某些人即使心理反對但不願或不敢表示，而只是以沈默不回答來因應，則整體的民意結果便會過份誇張該議題被眾人所接受的傾向（Berinsky, 1999）。由是，選樣偏誤模型的採用對於瞭解更真實具代表性的民意，應該有相當的意義。最後，期盼本文的提出能發揮拋磚引玉的效果，激發學界對於選樣偏誤問題的重視。

*　　*　　*

投稿日期：93.10.22；修改日期：94.03.13；接受日期：94.04.26。

附錄：資料來源與變數建構

本研究所使用的資料見於下表，皆係國立政治大學選舉研究中心所蒐集的電話訪問資料。

研究資料來源一覽表

選舉類型	計畫名稱	計畫主持人	執行日期	選舉日期	樣本數
2000 年總統	跨世紀總統選舉中選民投票行爲科際整合研究	陳義彥	2000/03/12 | 2000/03/16	2000/03/18	1582
2001 年台北縣長	新世紀台灣地區選民投票行爲之研究：民國 90 年台北縣長選舉投票行爲之研究	黃德福	2001/11/25 | 2001/11/30	2001/12/01	1802
2002 年台北市長高雄市長	台灣地區「分立政府」與「一致政府」之研究—民眾政治態度與分立政府之間的關連性：台灣經驗的探索(1/2)	游清鑫	2002/11/28 | 2002/12/06	2002/12/07	台北 1083 高雄 1081
2004 年總統	二〇〇四年總統選舉電話調查研究	劉義周	2004/03/06 | 2004/03/19	2004/03/20	1108

說明：表中所列研究計畫皆爲政治大學選舉研究中心執行的電話訪問調查。

本研究所使用的變數，及其測量方式如下說明：

一、選樣方程式

1. 表態與否：在選樣方程式中，應變數是表態與否，如果受訪者很明確地告訴訪員他可能的投票對象，則視爲表態，歸類爲 1；若受訪者的答案是不一定、尚未決

定等模擬兩可的答案時，視爲未表態，歸類爲 0；此外，若受訪者回答他不會去投票或投廢票，則歸類爲遺漏值。

2. 性別：性別區分男性與女性兩類，由訪員依照戶中抽樣對象自行輸入，因此不會有遺漏值的存在。
3. 年齡：詢問受訪者的出生年次，再加以換算而得。年齡拒答者歸爲遺漏值。
4. 教育程度：教育程度是詢問受訪者的最高學歷，然後將「小學及以下」、「國初中」合併爲「低教育程度」，「高中職」爲「中教育程度」，「專科」、「大學及以上」合併爲「高教育程度」。
5. 媒體使用：媒體使用情形變數所依據的題目是：「請問您在這次選舉當中，最常看那一份報紙的選舉新聞？」以及「請問您最常看哪一台電視的選舉新聞？」筆將受訪者的答案重新過錄爲「都不看」、「看其中之一」、「兩者都看」三類，依序給予 0 、1 、 2。
6. 政黨認同：此一變數主要是依照「政黨認同」的題目進行過錄。所依據的問卷題目爲：「在國民黨、民進黨、新黨、親民黨以及台聯這五個政黨中，請問您認爲您比較支持哪一個政黨？」，如果受訪者回答支持 XX 政黨，則續問「請問你支持 XX 黨的程度是非常支持還是普普通通」，如果受訪者在前一個題目的回答是沒有比較支持某一個政黨，則續問「請問您比較偏向國民黨、偏向民進黨、偏向新黨、偏向親民黨、偏向台聯，還是都不偏？」依照這三個問題可以將選民區分成非常支持某一政黨，普通支持某一政黨、偏某一政黨以及中立無反應等十六類。經過重新處理之後，支持國民黨、親民黨、新黨歸爲泛藍認同者，支持民進黨、台聯的歸爲泛綠認同者，其它的則歸爲中立無反應。
7. 候選人認知與評價回答情形：由於問卷題目的關係，這個變數在不同的選舉中，有不同的測量方式。2000 年總統選舉所依據的問卷題目爲「就您個人而言，您喜歡還是不喜歡連戰（宋楚瑜、陳水扁）？」連戰、宋楚瑜、陳水扁以隨機的順序出現。然後就受訪者對這三個題目的回答進行過錄，每回答一題給 1 分，如果三個問題都沒有回答，則爲 0，因此最高爲 3 分，表示受訪者對三個候選人都表示意見，0 表示受訪者對三個候選人都沒有表示意見。

在 2001 年台北縣長選舉中，問卷題目爲「在台北縣兩個主要的縣長候選人當中，如果我們用 0 分到 10 分來做標準，0 分表示您對他的印象很不好，10 分表示對他的印象很好。請問您會給蘇貞昌幾分？請問您會給王建煊幾分？」在這兩題中，如果兩題都沒有回答具體的分數，則歸爲 0，若其中一題有具體分數，則歸爲 1，兩題都回答具體分數則歸爲 2。

在2002年台北市長選舉的問卷題目中，則是先詢問受訪者「請問您認爲目前台北市最需要解決的問題是什麼？」再詢問他「請問您覺得馬英九或是李應元兩位市長候選人當中，哪一位較有能力解決這個問題？」如果受訪者在第二題回答的答案是「拒答」、「看情形」、「無意見」、「不知道」等無反應的答案，則歸爲「沒有回答」，其他則歸爲「有回答」。

而2002年高雄市長選舉的問卷上，有關候選人評價的題目則爲「請問在施明德、張博雅、黃天生、黃俊英、謝長廷這五位市長候選人當中，您最喜歡哪一位？」如果受訪者回答無反應的答案，則歸爲「沒有回答」，其他則歸爲「有回答」。

2004年總統選舉的問卷題目爲「接下來，我們想請教您對幾個政治人物的感覺。如果以0到10來表示，0表示非常不喜歡，10表示非常喜歡，5表示普通，請問0到10，您會給陳水扁（連戰、宋楚瑜、呂秀蓮）多少？」變數建構方式與2001年台北縣長相同，只是因爲在該次選舉中，不只總統候選人受到矚目，就連副總統候選人也備受關注，可說是相當的重要，因此筆者將對副總統候選人的評價也加入一起考慮，所以兩組候選人共有四道題目，變數的範圍是0到4。

8. 政治功效意識：政治功效意識這個變數是由兩個題目所組成的，僅在2002年的北高市長選舉中有這幾個題目，以台北市爲例，分別是「請問您覺得我們台北市民對台北市政府的施政有沒有影響力？」「請問您覺得台北市政府官員會不會重視我們台北市民的想法？」，筆者將這二個題目重新過錄，在第一題回答「有點影響力」、「非常有影響力」歸爲1，回答「不太有影響力」「非常沒有影響力」及「無反應」合併爲0，第二題回答「有點重視」、「非常重視」合併爲1，回答「不太重視」、「非常不重視」及「無反應」合併爲0，再將這兩題整理成一個政治功效意識的指標，從0到2來表示受訪者政治功效意識的高低，數字愈高，表示政治功效意識愈高。

二、結果方程式（outcome equation）

1. 投票對象：在投票對象變數建構方面，是詢問受訪者的可能投票對象，問卷題目在五次選舉的問卷中略有差異，2000年總統選舉、2002年北高市長選舉及2004年總統選舉的問卷都採取兩輪的問法，先詢問受訪者可能的投票對象，以2004年總統選舉爲例，問卷題目爲「如果明天就要選總統，在陳水扁、呂秀蓮和連戰、宋楚瑜這兩組參選人當中，請問您會支持哪一組?」，若受訪者的答案是尚未決定、無意見、不知道等無反應的選項，則再繼續追問可能偏向哪一位候選人：「那

您比較可能投給哪一組？」（註四）依照這兩個題目可以將選民可能的投票對象歸類出來。至於 2001 年台北縣長選舉則僅問一題：「如果明天就是投票日，請問：在蘇貞昌、王建煊、劭建興、石翊靖等四位候選人當中，您會投給誰？」由於2000年總統選舉、2001年台北縣長選舉、2002年高雄市長選舉的候選人當中，有幾組候選人的支持率與主要候選人相差太多，爲了避免樣本數過少而造成估計不穩定，因此在2001年的台北縣長選舉僅針對蘇貞昌與王建煊兩位候選人進行分析，2002 年高雄市長選舉，也是僅針對謝長廷與黃俊英進行研究，在 2000 年總統選舉中僅保留三組主要候選人（連蕭、陳呂與宋張），並且由於研究方法上的限制，應變數只適用於兩個類別，因此本研究將2000年總統選舉的投票對象變數改成投給陳水扁／不投給陳水扁的方式進行分析。

2. 候選人評價：候選人評價這個變數的建構方式因爲受到資料的限制，因此在不同選舉有不同的處理方式，而每個選舉所依據的問卷題目與選樣方程式中，候選人評價問題回答情形是一樣的，以下分別說明每個選舉的處理方式。2000 年總統選舉中，由於有三位主要候選人，因此在候選人評價變數建構上較爲特別，筆者比較受訪者對於三位主要候選人的喜歡程度之後，將之區分成七類，分別是「對連戰評價最高」、「對宋楚瑜評價最高」、「對陳水扁評價最高」、「宋扁同，對連戰評價最低」、「連扁同，對宋楚瑜評價最低」、「連宋同，對陳水扁評價最低」以及「無法比較高低」。

在2001年台北縣長選舉中，將受訪者對於兩位候選人的評價分數加以比較，如果給王建煊的分數高於給蘇貞昌的分數，則歸爲「對王建煊評價較高」，反之則歸爲「對蘇貞昌評價較高」，給兩者相同的分數，或是無反應的答案，則歸爲「無法比較高低」。

2002 年的台北市長選舉中，如果受訪者回答李應元較有能力解決他認爲的重要問題，就歸爲「對李應元評價較高」，反之則歸爲「對馬英九評價較高」，其他則歸爲「無法比較高低」。2002 年高雄市選舉，受訪者若回答較喜歡謝長廷，則歸爲「對謝長廷評價較高」，若較喜歡黃俊英，則歸爲「對黃俊英評價較高」，如果受訪者回答的是其他三位候選人，則歸爲遺漏值，其他答案則歸爲「無法比較高低」。

在2004年總統選舉中，有兩組候選人出馬競選，分別是民進黨的陳水扁、呂秀蓮，泛藍陣營的連戰、宋楚瑜。陳水扁在選擇副手人選時，曾經受到外界眾多的討論，最後選擇原有的搭檔呂秀蓮；而宋楚瑜是上屆總統候選人，以親民黨黨主席的身份擔任連戰的副手，營造泛藍團結的氣氛，因此筆者認爲，在該次的總

統選舉中，除了主要的總統候選人之外，選民對副總統候選人的評價，對於他的投票抉擇也具有相當的影響力，因此本研究將選民對於兩組候選人的評價加總起來之後，再加以比較，分成「對陳呂評價較高」、「對連宋評價較高」、「無法比較」三類。

3. 政黨認同：建構方式與選樣方程式中政黨認同變數的建構方式相同。
4. 對現任者施政滿意度：對現任者施政滿意度的變數建構方面，同樣因爲受限於資料而使用不同的問卷題目。2001 年台北縣長選舉的問卷題目是「整體來說，蘇貞昌做我們台北縣縣長以來的表現，0 分表示很不滿意，10 分表示很滿意，從 0 分到 10 分，請問您會給幾分？」2002 年北高市長選舉的問卷題目是「整體來說，您對馬英九（謝長廷）這三年多來擔任市長的表現滿不滿意？」2004 年總統選舉的題目型態也相當類似：「整體來說，請問您對陳水扁總統過去四年來的施政表現滿意不滿意？」本研究認爲，基本上，不管是用打分數的方法，或是回答滿不滿意的方式，都可以測得受訪者對於現任者的施政滿意度，只是採取的測量尺度有所不同。至於 2000 年總統選舉由於沒有測量現任者施政滿意度的題目，也沒有可供替代的題目，因此在該次選舉的投票模型中，將無法放入此一變數。
5. 統獨立場：建構統獨立場變數的問卷題目是「關於台灣與大陸的關係，有下面幾種不同的看法：⑴儘快統一；⑵儘快獨立；⑶維持現狀，以後走向統一；⑷維持現狀，以後走向獨立；⑸維持現狀，看情形再決定統一或獨立；⑹永遠維持現狀。請問您比較偏向哪一種看法？」爲了避免因選項類別過多而可能造成個別類別樣本數過少的情況，將⑴與⑶合併爲「傾向統一」，⑵與⑷合併爲「傾向獨立」，⑸與⑹合併爲「維持現狀」。
6. 省籍：建構省籍變數的問卷題目是「請問您的父親是本省客家人、本省閩南（河洛）人、大陸各省市人，還是原住民？」筆者將本省客家人、本省閩南人與原住民合併爲一類「本省人」，大陸各省市人歸爲一類「外省人」。

註　釋

註　一：在美國國家選舉研究（American National Election Study）中，評分的範圍是 0 到 100 分；而盛文中則是以 0 到 10 分爲評分範圍。

註　二：此處 Y^*_{2i} 是一個隱含（latent）變數，亦即理論上的變數，其與 X 呈現線性關係，是一個連續變數，而 Y_{2i} 表實際觀察到的變數，此處是二分變數，1 爲被選入，0 爲不被選入。通常此隱含變數若高於某一門檻（threshold，以 τ 表示），則 $Y_{2i}=1$，低於該門檻，則 $Y_{2i}=0$，但爲了使模型可求解（identified），因此通常令 $\tau = 0$，可參考 Long（1997：41, 122）的說明。

註　三：利用 GLS 校正標準誤的程序爲：

A. 將原有模型先使用 OLS 進行參數估計。

B. 接者針對每一個觀察值建立 Y 的預測值 p，將大於.99 的預測值過錄爲.99，小於 0.01 的預測值過錄爲 0.01，令 $q=1-p$，$s=\sqrt{pq}$。

C. 將迴歸模型中的自變數都除以 s，常數項爲 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爲加權最小平方法（weighted least squares, WLS）。

註　四：2000 年總統選舉問卷題目爲：「如果明天就要投票選總統，在連戰、陳水扁、宋楚瑜、許信良、李敖五組候選人當中，你會把票投給那一組？」，續問題目爲：「那您比較可能投給哪一組？」；2002 年台北市長選舉題目爲：「如果明天就是投票日，在馬英九與李應元兩個人當中，您會把票投給誰？」，續問題目爲：「那您比較可能偏向誰？」；2002 年高雄市長選舉題目爲：「如果明天就是投票日，在這五位候選人當中，您會把票投給誰？」，續問題目爲：「那您比較可能偏向誰？」。

參考書目

Ⅰ、中文部份：

何思因

1994 「台灣地區選民政黨偏好的變遷：1989-1992」，**選舉研究**，一卷一期：39-52。

李錦河和溫敏杰

1998 「從行銷學『產品屬性』角度建構『選民需求指標』選舉預測模式－以1997年台南市市長選舉爲例」，**選舉研究**，五卷二期：1-33。

洪永泰

1994 「選舉預測：一個以整體資料爲輔助工具的模型」，**選舉研究**，一卷一期：93-110。

范凌嘉

1999 「台灣縣市長選舉預測模型之研究：一個基礎模型的建立及其應用」，國立政治大學政治學系碩士論文。

徐火炎

1993 「選民的政黨政治價值取向、政黨認同與黨派投票抉擇：第二屆國大代表選舉選民的投票行爲分析」，**國家科學委員會研究彙刊：人文及社會科學**，三卷二期：144-166。

徐永明和林昌平

2003 「時序模型在選舉預測上的應用：以 Samplemiser 爲例」，「選舉預測模型」學術研討會，中央研究院中山人文社會科學研究所主辦。

張紘炬和丁台怡

2000 「1998高雄市市長選舉預測模型之比較」，**民意研究季刊**，二一二期：1-13。

張紘炬和林顯毓

1995 「台北市長選舉投票傾向的 LOGIT 模式分析」，**民意研究季刊**，一九二期：1-11。

張紘炬和黃男瑋

2000 「1998台北市市長選舉預測模型之比較」，**民意研究季刊**，二一一期：1-14。

梁世武

1994 「一九九四年台北市長選舉之預測：『候選人形象指標』預測模式之驗證」，

選舉研究，一卷二期：97-130。

1996 **選舉預測：一九九四年台北市長選舉中「候選人形象指標」預測模型之驗證**，台北：華泰書局。

盛杏湲

1998 「選民的投票決定與選舉預測」，**選舉研究**，五卷二期：37-75。

2002 「統獨議題與台灣選民的投票行爲：一九九〇年代的分析」，**選舉研究**，九卷一期：41-80。

盛治仁

2000 「總統選舉預測探討－以情感溫度計預測爲表態選民的應用」，**選舉研究**，七卷二期：75-108。

2003 「從立委和縣市長版圖預測總統選舉」，「選舉預測模型」學術研討會，中央研究院中山人文社會科學研究所主辦。

莊文忠

2000 「選舉預測與策略性投票：以八十九年總統選舉爲例」，**理論與政策**，十四卷二期：55-91。

陳陸輝

2002 「政治信任感與台灣地區選民投票行爲」，**選舉研究**，九卷二期：65-84。

陳義彥

1994 「我國選民的集群分析及其投票傾向的預測－民國八十一年立委選舉探討」，**選舉研究**，一卷一期：1-38。

游盈隆

1994 「台灣選民的議題投票－二屆國大選民的分析」，**東吳政治學報**，三期：219-254。

黃秀端

1997 「決定勝負的關鍵：候選人特質與能力在總統選舉中的重要性」，**選舉研究**，三卷一期：103-135。

溫敏杰、杜宜軒和李錦河

2000 「統計方法在選舉預測上之研究」，**民意研究季刊**，二一一期：40-65。

廖益興

1996 「影響選民支持李登輝與否的因素」，**選舉研究**，三卷二期：187-210。

廖達琪、景鴻鑫和楊連誠

2003 「類神經網路系統與選舉預測－二〇〇二年北高兩市市議員選舉之案例探

討」，「選舉預測模型」學術研討會，中央研究院中山人文社會科學研究所主辦。

劉子立

2004 「省籍族群政治與投票－台北市選民行爲之分析」，國立政治大學政治學系碩士論文。

劉文卿

1995 「台北市長選舉之基因預測模型」，**選舉研究**，二卷一期：1-16。

劉念夏

1996 「一九九六年總統大選選舉預測：民意調查中未表態選民投票行爲規則假設的提出與驗證」，**選舉研究**，三卷二期：131-156。

劉義周

1985 「調查研究中『不知道』選項問題之分析」，**國立政治大學學報**，五二期：65-90。

1996a 「選舉預測：一個簡單理論的試驗」，**選舉研究**，三卷二期：107-130。

1996b 「世代、統獨立場與投票抉擇：李登輝的選民」，「選舉制度、選舉行爲與台灣地區政治民主化」學術研討會，政治大學選舉研究中心主辦，1996 年 11 月 30 日、12 月 1 日。

鄭夙芬和陳陸輝

2000 「台灣地區民眾參與調查研究態度的變遷」，**選舉研究**，七卷一期：115-138。

謝復生、牛銘實和林慧萍

1995 「民國八十三年省市長選舉中之議題投票：理性抉擇理論之分析」，**選舉研究**，二卷一期：77-92。

Ⅱ、英文部份：

Achen, Christopher H.

1986 *The Statistical Analysis of Quasi-Experiments.* Berkeley: The University of California Press.

Almond, Gabriel A. and Sidney Verba

1963 *The Civic Culture.* Princeton, N.J.: Princeton University Press.

Asher, Herbert B.

1988 *Presidential Elections and American Politics: Votes, Candidates, and Campaigns Since 1952.* Belmont: Wadsworth.

Berinsky, Adam J.

1999 "The Two Faces of Public Opinion." *American Journal of Political Science,* 43: 1209-1230.

Brehm, John

1993 *The Phantom Respondents.* Ann Arbor: The University of Michigan Press.

Campbell, Angus, Gerald Gurin and Warren Miller

1954 *The Voter Decides.* Evanston, Ill.: Row, Peterson.

Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes

1960 *The American Voter.* New York: John Wiley and Sons, Inc.

Dubin, Jeffrey A. and Douglas Rivers

1989 "Selection Bias in Linear Regression, Logit and Probit Models." *Sociological Methods and Research,* 18: 360-390.

Fiorina, Morris P.

1981 *Retrospective Voting in the American National Elections.* New Haven and London: Yale University Press.

Heckman, James J.

1979 "Sample Selection Bias as a Specification Errors." *Econometrica,* 47: 153-161.

Long, J. Scott

1997 *Regression Models for Categorical and Limited Dependent Variables.* London: Sage Publications.

Rahn, Wendy M., John H. Aldrich, Eugene Borgida and John L. Sullivan

1993 "A Social-Cognitive Model of Candidate Appraisal." In Niemi, Richard G. and Herbert F. Wiesberg(eds.). *Controversies in Voting Behavior.* 3rd edition. Washington D. C.: Congressional Quarterly Press.

Sheng, Shing-Yuan

1994 "Seleciton Bias in Vote Choice Models." **選舉研究**，一卷二期：221-250。

Verba, Sidney

1996 "The Citizens as Respondent: Sample Surveys and American Democracy." *American Political Science Review,* 90: 1-7.

Selection Bias Models on Election Prediction

Ying-lung Chou* • Shing-yuan Sheng**

Abstract

When predicting elections, researchers always face the difficulty that some respondents do not answer the question of vote choice. However, including only those who give a clear answer on vote choice, researchers might have the problem of selection bias. This article tries to assess the effects of selection bias on vote choice model, correct the wrong estimates resulting from the bias, and predict elections accurately. In this article, we apply a bivariate selection bias model-- developed by Dubin and Rivers-- to five different elections. The research findings show that if researchers include only those respondents who give a clear answer on vote choice question, they might take a risk of overestimating the effects of independent variables on vote choice. This is because respondents who give a clear answer on the question of vote choice may also have definite and strong political preferences, and they are quite different from those who do not give a clear answer. After correcting the estimating errors resulting from selection bias, we might predict election outcome accurately. The largest predicting error in four elections is 1.16%. It is less than sampling error. The model cannot do a better job in only one election. Fortunately, it does not do a worse job, either. The overall result shows that the selection bias model is a reliable tool in predicting elections.

Keywords: election prediction, selection bias model, nonresponse in survey

* Ph. D. student in the Department of Political Science, National Chengchi University.
** Professor, Department of Political Science, National Chengchi University.

審查意見

審委意見（一）

這是一篇相當優秀的論文，探討選樣偏差對於選舉預測的影響，以及可能的矯正方法，在理論與實用上皆有貢獻，值得刊登，以下是幾點意見：

一、作者透過選樣偏誤模型來進行矯正是一個新的作法，傳統上都是以「加權」的方式進行，感覺邏輯是不同層次的，加權是針對成功樣本與母體間的代表性差異進行調整，而選樣偏誤模型則是以表態／不表態的差異來進行矯正，兩者的差別爲何並不一定是作者需要討論的，但卻是選舉預測準確的兩個前提，作者處理的資料是否有加權過，加權是否是作者模型的另一次矯正，可以略加討論。

二、另，在作者的模型中，是否已經篩選過不投票的觀察值，以願意投票的爲對象，以及如何確定投不投票的測量問題，不會干擾到選樣偏誤模型的測量。

三、在理論深化上，作者可以進一步探討表不表態的偏誤對於民調資訊的影響，如果選民表態有一定的政治傾向，當然會透過民調來放大其政治影響，這個趨勢在作者的研究中似乎是顯著，五個模型中有四次（2001 年台北縣除外），泛綠支持度在電訪表態、傳統 probit 模型中都被低估了，原因爲何？除了選民表態的傾向外，有沒有民調工具本身的偏差，或是比較大的政治結構因素。

審委意見（二）

本文應用 Dubin and Rivers 的 bivariate selection bias model，探究民調中選樣偏誤的問題及其解決之道。作者對國內選舉預測研究的文獻詳實，對選樣偏誤及其校正法的解說也相當詳盡，特別從 Heckman 到 Achen 再到 Dubin and Rivers 模型的論述非常仔細，對相關領域貢獻良多，是一篇可以刊登的佳作，筆者唯有下列幾點方法上問題供作者參考：

一、誠如作者在第一段所言，選樣偏誤的校正應立基於受訪者有系統性、不隨機的回應或不回應，如果抽樣的受訪者表態與否是隨機的，則對選樣所做校正的正當性便不夠強。然作者僅在該段第二節陳述過去所發現台灣選民不表態的情形，這樣的立論基礎似嫌薄弱，也未證實選樣確實是非隨機的不表態，所以建議作者在此引註多點文獻來佐證。

二、本文的焦點是應用 Dubin and Rivers 的模型到五次台灣選舉分析上，其中有幾點疑

惑煩請作者回應：

一、各選舉分析中，作者選樣方程式均有所差異，也就是在校正選樣偏誤上，使用的自變數隨不同選舉而有差別，其理由安在？此外，除了解釋不同選舉間的差異外，作者也應回應使用這些變項來做樣本校正的理論爲何？

二、Dubin and Rivers 模型中的兩組方程式各需 X_{1i} 與 X_{2i} 來做估算，但事實上兩組變項並不需完全排它（exclusive），它們可依各自的理論基礎加入同樣的解釋變數，所以 X_{1i} 與 X_{2i} 其中的變數是可以重複的。本文作者卻將兩方程式的自變項拆成不重複的兩組，也使得表 2 到表 6 中傳統 probit 模型必須跟著刪去多個解釋變數，若與表 1 模型相比，同樣傳統 probit 模型卻有不同解釋變數不免顯得唐突，所以建議作者可以考慮增加結果方程式的變項數，並統一傳統未校正選樣偏誤 probit 模型的自變數。

三、Dubin and Rivers 模型中的ρ係數相當關鍵，是詮釋不校正選樣偏誤的話是高估或低估結果方程式。本文分析的結果是 2002 台北市選舉影響不顯著，2001 年台北縣是高估民進黨候選人，其餘則是符合低估民進黨候選人的假設，這樣的結果似乎仍不盡滿意，作者雖在頁 35 多所解釋，但筆者認爲這也佐證了上述修改結果程式之議。

最後仍是整個模型設定的問題，表 2 到表 6 的 LR Test 是整個選樣偏誤模型的吻合度檢定，但除了兩次總統大選的投票模式估測外，其餘的模型均不符 $p < .05$ 的保守情況，或也顯示模型的確可以再加強。

審查意見答覆

審委意見（一）：

一、謝謝審查委員的提醒，實則本文中的模型皆未經過加權處理，藉以保持資料的原始樣貌。

二、在本文使用的模型中，已經篩選掉不投票的觀察值。至於在五個模型中有四次（2001 年台北縣長選舉除外），泛綠支持者在電訪表態、傳統 probit 模型中都被低估了，筆者已於文中說明，這與我們一般的認識相同，亦即民意調查的結果通常低估民進黨的得票率，至於 2001 年台北縣長選舉，也許是因爲在該次選舉裡，由於民進黨候選人爲現任縣長蘇貞昌，相當具有選戰的優勢，而泛藍陣營的候選人王建煊是在選舉日前不久才確定，而在整個選舉的過程裡，聲勢較弱，因此傾向投票給王建煊的受訪者比較不願表態。

審委意見（二）：

一、作者認爲如果受訪者是否願意表態是隨機的，則不致造成選樣偏誤的情形，然而，如果受訪者是否願意表態是系統性的，則會造成偏誤，而這也是本文要處理的問題。本文已在第四部份以及表 1 證實，受訪者的表態與否確實是系統性的：愈能夠區別候選人的差別，以及有政黨認同的受訪者，愈可能表態；同時，使用媒體愈多的愈傾向表態，政治功效意識愈高者愈傾向表態、男性比女性傾向表態，教育程度高者較傾向表態，而這些表態的都比較傾向於藍營的候選人。

二、㈠在進行選樣偏誤的校正之前，作者先建立選樣方程式的 probit 模型（如表 1），由於所使用的是二手資料，所以在模型估計時必須遷就於現有的資料，雖然在每一次的估計模型使用的變數並不相同，不過，這也正好可以證實，選樣偏誤模型相當具有穩定性，即便使用不完全相同的變數，但是校正的效果一樣都不錯。

㈡誠如審查委員所說，Dubin and Rivers 模型中的兩組方程式變數並不需完全排他，可依各自的理論基礎加入同樣的解釋變數，而在本文的模型中，筆者並未將兩組方程式拆成不重複的兩組。請注意表 1 所呈現的是選樣方程式（亦即估計是哪些因素影響受訪者表態或不表態），至於表 2 到表 6 中所呈現的，在傳統 probit 模型僅考慮選樣偏誤模型當中的結果方程式（亦即估計是哪些因素影響受訪者投票給哪一組候選人），而選樣偏誤模型中則同時考慮選樣方程式與結果方程式。

㈢ρ值只是反映出選樣方程式與結果方程式的誤差項的相關程度，如果相關程度高，則表示選樣偏誤嚴重。而如果 LR Test 無法拒絕ρ值等於 0，也只是顯示了選樣偏誤的情形不嚴重，此與模型估計的好壞無關。而模型究竟是否好，可以就各變數的估計參數是否符合理論上的預期爲判斷的標準，同時我們可以根據選樣偏誤模型（校正模型）與傳統 probit 模型（未校正模型）兩個模型的實際參數估計值，計算出每一個受訪者投給兩組候選人的機率，從而據以推論其投票對象，然後根據估計出的候選人得票比例與實際的選舉投票結果作比較，就可以知道校正模型是不是優於未校正模型。而在本研究中，選樣偏誤的預測結果在五次中有四次是相當準確的，在一次選舉預測中即使並沒有達到很好的結果，但也不會比不校正來得差。