# 8 項目無反應資料的多重插補分析*

劉正山**、莊文忠***

目次



* 論文初稿發表於臺灣選舉與民主化規劃與推動委員會主辦之「臺灣選舉與民主化調查2012年國際學術研討會：『成熟中的臺灣民主：TEDS2012調查資料的分析』」，2012年11月3-4日。本章資料來自行政院國家科學委員會補助之研究計畫，計畫編號為NSC 100-2410-H-128-001-MY2。作者感謝丘偉國、劉詩芃在資料處理上的協助。

** 國立中山大學政治學研究所副教授

*** 世新大學行政管理學系副教授

# 壹、前言

在社會科學領域，大多數的調查研究都是以「人」為對象，因此在資料蒐集過程中，即使是隨機化的抽樣設計和標準化的執行過程，仍然免不了會出現遺漏資料 (missing data) 的情形。換言之，受到調查執行過程中諸多主、客觀因素的影響，例如訪員找不到受訪者、受訪者不願意回答、資料處理出現瑕疵等等，以致調查資料多半是非完整的資料矩陣，亦即某些個案出現單位無反應 (unit nonresponse) 的情形。事實上，廣義的遺漏資料尚包括訪員漏問問題、問卷中跳題設計、受訪者因為對某些問題的資訊不足、過於敏感等而回答不知道、無意見、很難說等項目無反應 (item nonresponse) 在內。[1]在一般的情形下，多數研究者均是利用大型調查訪問成功的資料，經過加權處理單位無反應所可能引發的推論偏差後後即用於理論或假設的檢證。然而，即使克服了樣本結構偏差的問題，仍可能因為某些個案在描述統計上出現項目無反應而在分析時被排除，造成描述偏誤的情形。誠如 Filion(1976, 482) 所言，無論是採取哪一種資料蒐集方法，在大樣本的調查中，幾乎不可能取得每一位受訪者的答案；即使是再好的抽樣設計，樣本都無法代表母體；若樣本再產生資料缺失則會更容易導致推論上和預測上的誤導。不可諱言地，目前國內的面訪或電訪調查工作都面臨了這個方法上險峻的挑戰。

檢視相關文獻可以得知，學者在面對統計分析時因資訊遺漏所造成的結果偏誤（亦即理論上應該顯著不為零的變數卻因為遺漏值太多而導致其迴歸係數變得不顯著）時所使用的補救方法，大多是以插補法 (imputation) 為主，其中又以多重插補法 (multiple imputation) 為近年最受重視的插補法之一。不少研究已明確指出，使用多重插補法並**不會**破壞資料，或使迴歸

[1] 在社會科學調查中是以人為研究對象，經常會出現受訪者因參與調查意願不高、問卷題目過於敏感或缺乏與問題相關之資訊不足等因素而出現無反應的情形，Lohr(2010, 329)指出，在農業或野生生物的調查中，通常會以「遺漏資料」(missing data)這個名詞來取代「無反應」(nonresponse)，不過，兩者在概念上和補救方法上則是相似的。因此，本章交互使用這兩個名詞，並未特別區辨這兩個名詞的異同。

分析的結果更差（見 Liu, 2010 的歸納）。這有助於我們進一步推判，除了把多重插補法用於推論統計之外，多重插補法也有應用於描述統計的潛力，這對於政治學的經驗研究中類別型資料分析及增進描述統計精準度的提升或許有幫忙，只是此一方法能有多大程度提升資料在描述統計上的品質？目前政治學界尚無足夠的研究成果可供歸納出具體結論。

因之，本研究使用數筆國內大型面訪資料庫「臺灣選舉與民主化調查 (TEDS)」的資料（2004、2008、2012 年總統選舉），針對樣本中有去投票的受訪者，以多重插補法重新計算樣本中有去投票的受訪者他們的投票選擇和分配，比較使用多重插補與不插補 ( 單純使用加權過後的百分比 ) 的差異，並與實際選舉得票結果作為對照。本研究預期得到兩個面向的結果：使用多重插補法用於描述統計上多大程度增加描述的精準度，以及學界是否可以在未來的研究中使用並信任多重插補法對於變數的描述，這些結論將有助於學界進一步認識、評估多重插補法的應用。以下在第貳節將就遺漏資料的定義及目前的主要處理方法作一回顧；在第參節詳細說明目前針對遺漏資料的插補方法；我們根據文獻中所整理出來的主流方法「多重插補法」進行研究設計，在第肆節說明所使用的資料及工具；第伍節呈現分析結果，並在最後一節進行討論。

## 貳、遺漏資料的意涵與處理

遺漏資料或不完整資料 (incomplete data) 可以說是量化研究中最常見及最容易干擾分析結果，卻也是不好處理的問題。Little 和 Rubin(2002) 將遺漏資料界定為「在現實世界中真實存在，但因為設備或人為等因素而無法確實得到資料」。陳信木與林佳瑩（1997）指出，由於不適當的研究設計或資料蒐集過程所造成的疏失，可能導致樣本中有部分資料或全部資料出現缺漏的情形。一般而言，調查資料中所出現的遺漏資料或是不完整資料，可以細分為兩大類型 (Berinsky 2008; Dillman et al. 2002; 陳信木與林佳

瑩 1997；杜素豪 2004)：「單位無反應」(unit nonresponse) 和「項目無反應」(item nonresponse)。單位無反應亦稱之為「個案無反應」，係是指受訪者根本沒有接受訪問，所有樣本中有部分觀察體之全部資訊遺失；項目無反應則是受訪者雖然接受了訪問，但卻在部分問題上沒有提供研究者有效的分析答案（如回答不知道、無意見、很難說或拒答），導致樣本中某些觀察體的部分（變項）資訊遺失。邱皓政（2003, 2）對此指出，遺漏狀況的最大影響是樣本的流失，造成研究資源的浪費或樣本不足的問題，但是，如果資料的遺漏具有特定的趨勢或傾向，此一系統性遺漏對分析結果將造成更嚴重的影響，必須小心處理及予以補救。

易言之，遺漏資料是大多數研究者所不樂見的事實之一，它們常會破壞研究計畫的假設，直接威脅研究的內在效度（統計的檢定力）及外在效度（類推結果至標的母群體的程度），即使研究者能以適當的策略來處理遺漏值，不同的處理方法也可能導致不同的結論；若忽略遺失值逕行予以刪除，容易使剩餘的樣本失去代表性，除了影響分析結果外，也可能造成結果的偏誤。誠如廖培珊等（2011, 147）所指出，除非無反應是屬於「隨機」發生，否則，刪除遺漏資料不僅導致可用以分析的樣本數減少、提高非抽樣誤差，還會影響到分析結果的通則化程度。是以，對於遺失值採取適當的處理方式，除了能有效使用含遺失值的資料外，分析上亦可減少錯誤結論的風險（鄒慧英與江培銘 2012, 3）。

整體而言，Lohr(2010, 330) 指出，在資料蒐集過程中，處理遺漏資料的四種途徑及其特性如下：(1) 利用調查設計減少無反應比例，以預防無反應的發生，這個方法很明顯是最好方法，只是研究者必須在調查規劃階段投入更多的資源進行周全的安排。(2) 設法取得無反應者的有代表性次樣本，利用這些次樣本來推論其他的無反應者，例如投入部分資源追蹤失敗樣本，以取得其相關的資訊。(3) 利用模型來預測無反應者的數值，例如加權調整方法隱含利用一個模型來調整單位無反應，插補方法通常是用來調整項目無反應，而參數模型可能被用來處理這兩種無反應的類型。(4) 忽略無反應的個案，雖然 Lohr 並不建議採取此一方法，但很不幸地，這是實務上最常見的方法。事實上，如果遺漏資料的比率甚低，對分析結果

的影響不大，可以忽略或不處理這些遺漏資料；反之，若遺漏資料的程度是不可忽略的，那麼，以這些受訪者為基礎所做的推論可能就會發生嚴重的錯誤。要言之，受訪者不回答問題除了影響樣本代表性之外，當未回應者與有回應者之間具有系統性的差異時，若單以有回應者的答案對母體進行推論，便會潛存未回應的偏差 (nonresponse bias)。

Hair 等 (2010, 46-50) 則是從資料分析的觀點提出遺漏資料的處理步驟：(1) 決定遺漏值的型態，判斷資料的遺漏是可忽略的 (ignorable) 或不可忽略的 (not ignorable)，如係前者，遺漏的資料僅是造成樣本數減少，對分析結果的影響不大，如係後者，便進入下一步驟；[2](2) 決定遺漏值的嚴重程度，如果遺漏資料集中在一小撮個案或變數，可將個案或變數刪除，對分析結果的影響不大，如果遺漏資料較為分散，便進入下一步驟；(3) 診斷遺漏值的隨機性，確認資料是完全隨機遺漏值 (Missing Completely at Random, MCAR)、隨機性遺漏值 (Missing at Random, MAR) 或非隨機遺漏值 (not missing at random, NMAR)，如係前二者，採取何種補救方法的效果差異不大，如係後者，便進入下一步驟；(4) 選擇適當的插補方法，利用樣本中其他變數／個案的有效值的基礎上，估計遺漏值的過程，目的是運用樣本的有效值中的已知關係協助估計遺漏值。由此可知，分析資料發生遺漏的模式 (pattern)，其重要性並不亞於計算資料遺漏的數量 (amount)，可以幫助研究者瞭解資料遺漏的可能機制與影響，再決定是否採取嚴謹的估計程序，以對症下藥（邱皓政　2003, 2.11）。

研究者除了在調查事前的準備階段設計各種策略預防遺漏資料的發生外，如果調查執行工作已經完成，研究者除了刪除無反應或遺漏資料的個案外，另一個作法是先對遺漏資料進行補救，再以補救後的資料進行統計分析。由相關研究文獻可知，就不同資料遺漏類型的補救而言，單位無反

---

[2] 可忽略的(ignorable)遺漏資料包括：(1)因為自母體中抽取樣本而導致母體中某些個案未被蒐集資料；(2)因為問卷設計而產生的跳題回答；(3)因為調查尚未完成而造成的資料不完整。不可忽略的(not ignorable)遺漏資料包括包括：(1)已知的(known)：如資料蒐集與處理過程因人為因素所造成的；(2)未知的(unknown)：難以發掘或調整的遺漏資料，大多是和受訪者本身的回答有關，如拒答(Hair et al. 2010, 44)。

應的主要補救方法是採取加權調整的方式，以減少有反應者和無反應者之間的差異對整體估計誤差的影響，不過，由於此一方法是建立在有反應和無反應者之間是沒有差異的假定基礎上，所以依然會存在有偏差的估計 (Filion 1976, 484)。而項目無反應是指受訪者雖然採取合作態度，願意接受訪問，但對某些問題並未提供有意義的答案，基於實務上（不必然是統計上）的理由，通常是採取插補的方式來補救這些遺漏資料 (Chen and Shao 1999; Mason, Lesser, and Traugott 2002; 陳信木與林佳瑩　1997)，一旦完成此一步驟後，插補的資料就會被視為是觀察的資料，研究者便開始進行後續的統計分析工作。

綜言之，研究者必須特別關注如何處理遺漏資料的議題。早期對遺漏資料的處理並無一特定的方式，大多由研究者或根據主觀認知，依照題目的類型或由所欲考察之內容，來決定處理這些類屬的方式（伊慶春與蘇碩斌　1995）。例如就項目無反應而言，在多數研究中，最常見的處理方式是將之視為是遺漏資料而予以捨棄。不過，近年來利用插補方法來處理迴歸分析中的項目無反應或某些不完整資料形式已是很普遍的策略，因為研究者可以將資料檔中的「漏洞」填滿，有助於利用完整資料分析方法進行統計分析。然而，要小心的是，如果研究者沒有特別留意插補方法的合理性及其潛藏的風險，此一作法也可能導致「欺騙的」統計推論結果。誠如論者所言，插補的觀念是既誘人又危險的 (seductive and dangerous)，其誘人處在於它可能誘使使用者在相信資料是完整的愉悅狀態下失去戒心，其危險處在於其同時容忍兩種情境 (Hair et al. 2010, 50)：一是可以被利用插補方式正當處理的問題是完全不重要的情境；二是將標準估計值應用在真實值上及插補資料在本質上存有偏差的情境。Hair 等 (2010, 42) 強調，研究者的首要之務是找出隱藏在遺漏資料背後的模式和關係，再利用補救方法，才能確保接近原始資料的數值分布。此外，資料的遺漏程度也是非常重要的，因其會影響到資料插補方法的選擇。

# 參、遺漏資料的插補方法

顧名思義，所謂「插補」意指研究者依據某種規則或演算法，對資料中未作答的無反應項目給予一個或多個合理數值，取代該遺失值，以保存資料矩陣的完整性。Rubin(1987)、Little 與 Rubin(2002) 進一步指出，若僅將每筆遺漏的資料填補一個合理的數值，稱為「單一插補」(single imputation)，若將每筆遺漏的資料填補多個合理數值，則稱為「多重插補」(multiple imputation)。無論是哪一種作法，都可以讓研究者在不減少樣本數的前提下建立完整資料，其主要目的就是盡量降低遺失資料在參數估計上造成的誤差，藉由有效插補後的資料檔，可使分析結果趨於一致，甚至優於不完整資料的分析結果（鄒慧英與江培銘　2012, 8）。

廖培珊等（2011, 150）亦指出，將遺漏值之外的無反應選項與各題目的不同答案合併，或者以特定數值來取代的方式，均屬於簡易插補法，在遺漏值比例甚低或資訊不足的條件下，常被社會科學研究者所使用。然而，簡易插補法並未考慮樣本中觀察值在其他變項間的相關，容易造成變異量被低估，且鮮少以統計模型予以檢證，因此，在實證研究中已較少見。目前可見的各種插補方法大多是在資料分析階段利用輔助變數 (auxiliary variables) 進行統計演算來獲取遺漏資料的替代數值（許玉雪與林建弘　2008；陳信木與林佳瑩　1997；廖培珊等　2011）。Lohr (2010, 351) 指出，透過插補創造出「乾淨的」矩陣式資料檔後可供研究者利用一般的軟體進行分析，對資料的不同次集合進行分析也可以得到較合理的結果，尤其是當無反應是隨機發生時，插補實際上可以降低因為項目無反應而造成的偏差。以下就資料分析時常見的幾種處理方式及其優劣之處作歸納比較。[3]

---

[3] 下文所忽略未談的一種類型是「演繹插補」(deductive imputation)，它是一種在資料編輯階段即可插補的方法。此一方法利用變數之間的邏輯關係來插補遺漏資料，例如受訪者在「是否曾經是暴力犯罪的受害者？」這個題目上是遺漏資料，但在前一題目「是否曾經是犯罪的受害者？」回答沒有，所以，後一題的遺漏值可以合理修正為沒有；又例如在縱時性的研究中，如果有婦女在第一年的調查中回答目前有1個小孩，第二年的調查中是遺漏資料，第三年的調查中回答目前有2個小孩，則第二年的遺漏值在

## 一、完整個案分析(complete case analysis)

當研究者發現有遺漏資料時，最簡單的處理方法即是刪除這些資料，僅保留完整的資料納入分析（邱皓政 2003, 2.12），嚴格來說，此一作法並不是對遺漏資料進行插補，乃是假定所有的遺漏資料都是隨機產生，而利用擁有有效資料的個案來代表全部的樣本 (Hair et al. 2010, 51)。詳細言之，使用完整個案分析有兩種作法：第一種是將任有有遺漏值的觀察值整筆刪除 (list-wise deletion)，亦即當受訪者有任一變項出現遺漏值時，便將該受訪者的所有資料整筆刪除，不列入後續資料分析。此一策略的優點在於容易執行，刪除過後的資料可直接進行一般統計分析，且由於所有變數來自相同樣本，故可進行單變量統計的比較；其缺點一方面是可分析的樣本減少，使得剩餘的樣本不具代表性和導致統計考驗力下降，另一方面是易受到遺漏資料是非隨機性而導致系統性偏誤的產生 (Hair et al. 2010; Little and Rubin 2002; Peugh and Enders 2004; 陳信木與林佳瑩 1997；邱皓政 2003；鄒慧英與江培銘 2012)。因此，當遺失資料非屬完全隨機遺漏 (MCAR) 且遺失比率不算小時，整筆刪除就不適合使用。

第二種是成對刪除法 (pair-wise deletion)，指的是在進行統計分析時保留遺漏資料，等到分析時涉及到這些個案或變數時才將該遺漏資料刪除，故又稱之為「所有可用資料分析」(all available data analysis)。此方法的優點是儘可能極大化可以利用的資料，避免樣本的大量減少；其缺點是會產生不一致的相關係數矩陣，甚至此一相關係數矩陣的特徵值 (eigenvalues) 可能出現負值。由於這個方法會改變相關係數矩陣的變異性質，整個研究的檢定力也會隨之變動，導致在接下來利用多變量分析進行統計推論困難重重 (Hair et al. 2010; 陳信木與林佳瑩 1997；邱皓政 2003)。

無論是採取哪一種刪除方法，都會造成統計檢定力的下降，不過，由於此一方法具有簡單、不涉及複雜的統計理論，且大多數統計分析軟體工具在執行迴歸分析時即「預設」了這個方法，它在目前國內外社會科學的

---

邏輯上可以插入數值2(Lohr 2010, 347)。

研究中可說是主流的作法。此外，值得一提的是，此一處理方式雖然避免了使用不完整資料所可能產生的偏誤，但是那些被刪除之資料所潛藏的資訊則是隨著資料的刪除而完全被忽略了，尤其當資料屬於非隨機遺漏(NMAR) 時，亦即有某種特徵的受訪者容易產生遺漏資料時，變數之間的關係就有可能遭受到扭曲。

## 二、平均值插補(mean imputation)

這是目前使用插補方式來挽救遺漏資料時最常用的方法。此方法適用在遺漏值是隨機發生的前提下，以該變數有回應樣本所回答之答案的平均數來取代遺漏資料，其立論基礎是平均值為最佳的單一替代值，可以反映該變數的集中情形。此方法的優點是易於了解和執行，且可以保留完整資料進行分析，適合用在有遺漏資料之個案較少的情況下；其缺點是未能反映無反應者的變異性，會造成變異量的低估，使得變數之間的相關性減弱及扭曲觀察資料的數值分布 (Lohr 2010; Hair et al. 2010)。所以，如果欲插補的變數接近常態分布，則研究者可以在維持平均數和標準差不變的原則下採取隨機插補法，以減少變異量被低估的情形 (Lohr 2010, 348)。

此種插補方法的一種變體是「群體平均值插補」，即利用第二個變數將所有個案分為不同群體，依據受訪者所屬的群體，利用各群體內有效資料之個案計算所得之平均值來替代該群體內有遺漏資料的個案，更能精確反映該受訪者所屬族群的特性（邱皓政　2003, 2.14）。另一種變體是，如果遺漏資料是類別變數，便無法以平均值取代，可改以眾數替代 (mode imputation)，其缺點是眾數可能是不存在或不只一個。

## 三、熱卡/冷卡插補(hot/ cold deck imputation)

此一方法是利用其他變數的資訊找到適當數值來替代遺漏資料，分為熱卡插補和冷卡插補兩種作法。熱卡插補是依照輔助變數的不同條件，

將未出現遺漏值的個案分類成若干「插補細格」(imputation cell)，再將出現遺漏值的個案依其在輔助變數的條件，從相對應的插補細格中尋找相似特徵的個案，以其數值替代遺漏值（陳信木與林佳瑩　1997；邱皓政　2003）。換言之，熱卡插補是以變數的交叉分類為基礎來界定插補的格子，由若干個輔助變數所形成的眾多「插補空格」必須是彼此周延(exhaustive)、互斥 (exclusive)、和同質的 (homogeneous)。此方法的優點是成本低，可減少單變量統計的無反應偏差、降低僅使用部分完整資料進行分析所導致的變異（因可分析的樣本數增加）、插補值會落在實際觀察值的範圍之內、易於分析及跨使用者之間的一致性（因無論是 X 和 Y 的交叉表格或 X 和 Z 的交叉表格，都會得到相同的 X 平均數）等 (Marker et al. 2002, 329-330)。

同樣屬於熱卡插補法的另一種作法稱之為「最鄰近熱卡插補法」(nearest-neighbor hot-deck imputation)，即藉由衡量觀察個案之間的距離，以最接近遺漏資料之受訪者的數值來插補該遺漏資料的個案 (Lohr 2010, 349)。Fix 與 Hodges(1951) 提出「最鄰近法則 (nearest neighbor rule)」原理如下：給定一組包含 $n$ 個樣本 $(x_1, \theta_1),(x_2, \theta_2),...,(x_n, \theta_n)$ 的集合，其中 $x_i$ 為第 $i$ 筆資料的觀察值，$\theta_i$ 為第 $i$ 筆資料所屬類別；若有另一筆資料 $(x, \theta)$，其中 $x$ 是可觀察的，我們可利用 $n$ 筆已知訊息的資料將 $x$ 歸類（轉引自廖培珊等　2011, 151-152）。[4]

---

4 由於資料歸類時，涉及兩筆資料距離(distance)的遠近，或是相異度(dissimilarity)的大小，Kaufman與Rousseeuw(1990)提出一種適用於混合不同測量尺度資料的相異度定義；假定每一個觀察值由多個不同類型的變項所構成時，兩個觀察值之間的相異度定義為：

$$d(i,j)=\frac{\sum_{f=1}^{P}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{P}d_{ij}^{(f)}}$$

其中$d(i, j)$表示第$i$筆觀察值與第$j$筆觀察值之間的相異度，其範圍介於0到1之間。$d$的意義為在第$f$個變項之中，第$i$筆觀察值與第$j$筆觀察值間的相異度，$f$=1,2,...,$p$。依據變項類型的不同，$d$的計算方式包含三類：(1)變項f屬於二元(binary)或名目(nominal)資料；(2)變項$f$屬於區間(interval)資料；及(3)變項f屬於順序(ordinal)或比率(ratio)資料。利用其定

Marker 等 (2002, 230) 指出，熱卡插補法有三個使用上的顧慮：(1) 通常不會對每一個標的變數都達到最佳化，因為實際上，每個目標變數的最佳預測變數都是不一樣的；(2) 雖然每一個變數所插補的一組數值都是可能的，但當進行交叉分析時，這些數值可能是無意義的；(3) 以插補過的資料檔為基礎的單純變異估計的偏差會下降，導致信賴區過度的狹小。近年學界持續在探討新的熱卡插補法來改善單變數的 MSE 和避免關聯性的稀釋。[5]

相較於熱卡插補的作法，冷卡插補則是利用過去的調查資料，同樣依照輔助變數的不同條件分成若干插補細格，將出現遺漏值的個案依其在輔助變數的條件，從相對應的插補細格中尋找相似特徵的個案，以其數值替代。Hair 等 (2010, 53) 提醒，若是採取冷卡插補的作法，研究者必須確認以外部來源取得的替代值比內部產生的替代值更為有效。

## 四、個案替代(case substitution)

此一插補方法是將有遺漏資料的全部個案透過選擇其他非抽樣的個案取代，最常見的例子是被抽中的家戶無法接觸或有大量的遺漏資料時，便以另一個原本不在樣本內的家戶取代，此一替代的樣本最好和被替代的原始家戶相似。雖然此一方法可以用在替代有少量遺漏資料的個案，但最常用在替代完全遺漏資料的個案（即單位無反應）。此一方法有爭議之處在於，研究者是否具有取得不在樣本內之個案的能力 (Hair et al. 2010, 53)。

---

義即可計算出資料由不同類型變項所組成時，兩筆資料的相異度。此一方法應用多見於經濟學與心理學（轉引自廖培珊等　2011, 151-152）。

[5] 例如R套件中的“StatMatch”( http://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf)與“VIM”(http://cran.r-project.org/web/packages/VIM/index.html)等開發團隊都持續在此方面努力。

## 五、模型基礎法(model-based methods)

Hair 等 (2010, 50-51) 指出，如果遺漏資料是非隨機的 (MAR)，研究者僅能利用特殊設定的模型來插補，其他任何插補方法的應用都可能導致偏差。此方法的插補途徑有二，第一種途徑主要是利用最大概似法 (Maximum likelihood method) 的估計技術，先將遺漏資料模型化，再找出最精確和最合理的可能估計值，如 EM 模型即是一例，這是重複的兩階段方法（E 階段和 M 階段），E 階段是對遺漏資料做最佳的可能估計值，M 階段則是以假定遺漏資料已經被替代的參數（如平均數、標準差或相關）進行估計，重複地進行此一過程，直到估計值的變化是可以被忽略的為止，以最後的估計值來取代遺漏值。此途徑的優點是只要模型正確，即使遺漏值的分布並非完全隨機，最大概似法所求取的估計值仍然會是一致的 (consist) 和有效率的 (efficient) （陳信木與林佳瑩 1997）；其缺點是在執行上和解釋上較為複雜。第二種途徑是將遺漏資料直接納入分析，將有遺漏資料的個案界定為樣本的一個次集合，此一途徑最適合用在相依關係中的自變數上，除了將原始變數放入模型外，並利用虛擬變數的作法，將有效資料編碼為 1，遺漏資料編碼為 0，此一虛擬變數即代表有效資料的個案和遺漏資料的個案在依變數上的差異，如果係數達到統計上顯著，即表示此一差異是確實存在的，而原始變數的係數即代表無遺漏資料的所有個案的關係。

## 六、迴歸插補(regression imputation)

此一方法的原理是應用方程式 $Z=\beta_0+\sum_{k=1}^{k}\beta_k X_k+\varepsilon$ 來進行遺漏值插補，即利用有遺漏資料之變數與資料檔內其他變數的關係為基礎，以其他變數為預測變數，預測遺漏資料的數值 (Hair et al. 2010; 陳信木與林佳瑩 1997；邱皓政　2003)。詳言之，若全部樣本中只有 m 個觀察體的 Z 變項未流失，且假定 Z 變項與一組輔助變項 ($X_1$, $X_2$...$X_k$) 之間存在線性關係，

則研究者可以用一群完整個案的依變數和自變數建立迴歸分析模型（即以 m 個觀察體計算上述的迴歸方程式，估計其迴歸參數值），再將有遺漏資料的個案，依其在自變數的條件，利用此一迴歸模型預測其數值，以取代遺漏值。若需插補的依變數為類別變數，則可改用 logistic 迴歸分析、區辨函數分析 (discriminant function analysis) 等其他一般線性模型建立預測模式。

此方法的優點是考慮變數間的相關性，不同個案會因其自變數的條件不同而以不同的數值取代，而不是單純地以平均數或固定數值取代，估計的基礎更為豐富，精確度得以提高（邱皓政　2003）。然而，這個方法有幾點需要考量：(1) 強化資料檔內已經存在之關係，會導致資料更具有樣本的特徵而無法通則化；(2) 除非估計值加入隨機項，否則容易低估分布的變異性；(3) 此一方法假定有遺漏資料的變數和其他變數之間有實質關係存在，如果這些關係並不足以產生有意義的估計值，那麼，其他方法 ( 如平均值插補 ) 反而是較好的選擇；(4) 樣本數必須非常大，才有充分的觀察個案可用以建立此一預測；(5) 迴歸分析所產生的估計值並沒有範圍的限制，因此，若預測值沒有落在依變數的有效範圍內 ( 例如在十點制量表預測得到 11 的數值 ) 時則需要調整；(6) 迴歸分析的過程較為繁複，不同的變數出現遺漏值便需進行一次迴歸分析 (Hair et al. 2010; 陳信木與林佳瑩　1997；邱皓政　2003)。

## 七、多重插補(multiple imputation)

多重插補法是迴歸插補法的一種，也是模型基礎法的延伸，因為它是目前插補法中最受推崇的主流方法，因此我們將它獨列一項來介紹。多重插補法是由 Rubin 於 1978 年首先提出，再由 Little 和 Rubin 加以發展，後續有諸多學者投入此一方法的探討。多重插補是由單一插補法延伸而來，研究者選擇多個參考變數後、產生 m 個 (m > 2) 完整的插補資料檔。接下來研究者使用每個資料檔進行分析，再將這些因多個資料檔得到的多個參

數進行合併，得到參數的估計與標準差。換言之，研究者自一個預測值的合理分配（如多變量常態分配）中隨機抽取多個數值進行插補，再分別對每次插補結果進行分析，接著跨多次插補分析結果取平均值，來獲得所欲之估計參數與標準誤。此一方法可適用在時貫性資料及單一觀察值資料，也可處理多變量資料結構。

多重插補的另一個特色是可使用不同演算法來產生插補值，目前常見者為最大期望法 (expectation maximization, EM)[6]與馬可夫鏈蒙地卡羅法 (Markov Chain Monte Carlo, MCMC)。[7]EM 演算法為兩步驟的迭代法，每次迭代均包含 E 步驟 (expectation step) 與 M 步驟 (maximization step)，在缺乏可觀察資料或混合模型下，此法藉由反覆迭代程序計算出最大概似估計值（鄒慧英與江培銘　2012, 10）。蒙地卡羅法的概念是將遺漏的資料填補 m 個數值，連同原有的其他變項與數值產生 m 個資料檔，以此進行分

6 此法最早由Dempster、Laird與Rubin(1977)所提出，用於求取不完整資料的最大概似估計值。首先是給定參數估計值(起始值)的條件下，透過可觀察資料計算完整資料的條件期望值，此為E步驟，亦即將不完整資料或遺失資料用條件期望值替代；接著根據E 步驟所得的條件期望值，最大化完整資料的概似函數，以求取參數的最大概似估計值，即完整資料之最大概似估計，此為M步驟。透過不斷重複這兩步驟（反覆迭代），直到結果不再變動或收斂到設定的精確度，此結果即是最大概似估計值（鄒慧英、江培銘 2012, 10）。

7 MCMC策略為Tanner及Wong(1987)所提出的資料擴增法(data augmentation)，將收斂到穩定分配(stationary distribution)的兩階段插補步驟執行m次之後，即可完成資料插補（廖培珊等　2011, 151）。MCMC法是貝氏估計法(Bayesian inference)的延伸，貝氏定理旨在說明藉由加入新觀察值來即時更新先驗機率(prior probability)，以得到後驗機率(posterior probability)。換言之，當真實的資料未能符合統計上所預設的假設或其分配未知時，通常可以透過貝氏方法(Bayesian approach)整合先驗分配(prior distribution)或先驗機率(prior probability)來分析資料，不過，此方法在理論上雖然可行，但實務上要找出未知參數$\theta$的後驗分配卻十分困難，或是不易進行模擬（廖培珊等　2011, 151）。MCMC法在插補遺失值的應用上，其終極目標在於取得一個最接近完整資料的目標分配，再據此分配進行插補。理論上，MCMC法會先給定各個參數的起始值，並依據可觀察資料建立起條件後驗分配(conditional posterior distribution)，接著從條件後驗分配中進行反覆抽樣，進而建構出一長串的馬可夫鏈樣本，再藉由馬可夫鏈中的隨機變數數值求得後驗分配的近似分配，此迭代程序須反覆至近似分配收斂至目標分配時止，或近似分配的相關統計量達到收斂標準。理論上，抽樣樣本數愈大，所得到的參數估計值愈逼近於實際的後驗分配，一般實務作法係觀察抽樣結果，當模擬誤差小於樣本標準差的5%時，即可視為較佳的逼近值（鄒慧英與江培銘　2012, 10-11）。

析，插補後的參數估計值為 m 個資料檔所得之平均值，當插補次數達到 10 次之後，再增加插補次數其相對效率並不會提升太多，因此 m 通常必須大於 3，但是不需要超過 10(Rubin 1987; 廖培珊等　2011,151)。此一方法的優點是插補值是從多個模型中隨機重複抽取而得，且可比較不同模型的敏感性，可以提高估計值的有效性；其缺點是計算較為複雜，若無適當的軟體輔助則操作起來較不容易。

Hair 等 (2010, 54) 建議，如果有許多的插補法可用，研究者應該考慮採取多重插補的策略，即結合數種插補方法以取得複合估計值 (composite estimate) ——通常是各種插補法的平均值，來替代遺漏資料，其立論基礎是使用多重插補方法可以將任何單一插補的特殊考量極小化，且此一複合估計值將是最有可能的估計值。Lohr(2010, 351) 也指出，此一作法可以讓研究者對不同模型對特殊無反應模式的插補結果更有敏感性。由於後來的分析者往往無法區辨原始和插補的數值，因此 Lohr 建議使用插補方法者應該對其插補的過程和接受插補的個案做完整的紀錄。

綜論之，早期大部分的分析並未將遺漏資料納入考量，多數統計軟體的預設分析模式不管是刪除遺漏變數或完整個案分析，都有可能產生偏差的、無效率的和不一致的分析結果。近年所發展的插補方法係指用各種策略來「填補」調查中的項目遺漏資料，一般的插補方法可以用來分析完整的資料檔，而特殊的多重插補方可以用以取得反映插補不確定性的標準誤。因為簡易插補法的應用缺少理論依據，而平均數插補法則有變異量被低估的問題，因此目前多數處理到遺漏值的研究多半以多重插補或運用多個輔助變項的插補方法（例如迴歸插補或熱卡插補）來進行（陳信木與林佳瑩　1997）。事實上，各種插補方法各有其優缺點，沒有哪一種插補方法在各種情境中都是最佳的選擇，研究者應該仔細地檢視資料遺漏的情境，再從中選擇最適當的插補方法。從實務的觀點來看，對資料進行插補只是分析過程的步驟之一，若耗費過多成本來處理對研究者來說可能是一大負擔（廖培珊等　2011, 170）。茲將以上的資訊整理到表 8.1。

**表 8.1 遺漏資料的插補技術比較**

| 插補方法 | 優點 | 缺點 | 最佳使用時機 |
|---|---|---|---|
| 只利用有效資料的插補 | | | |
| 完整個案分析 | • 最容易執行<br>• 許多統計軟體的預設方法 | • 最容易受到非隨機過程的影響<br>• 樣本數的損失最多<br>• 較低的統計檢定力 | • 較大的樣本數<br>• 變數之間有較強的關係<br>• 資料的遺漏程度較低 |
| 所有可用資料分析 | • 有效資料的最大利用<br>• 在不替代數值的之下儘可能地將樣本數極大化 | • 每一個變數插補的樣本數不一樣<br>• 在相關和特徵值的計算可能產生「超出範圍」的數值 | • 資料的遺漏程度相對較低<br>• 變數之間是中等相關 |
| 利用已知的替代值插補 | | | |
| 個案替代 | • 提供真實的替代數值而不是計算得到的數值（例如另一個實際的觀察值） | • 必須有不在原始樣本內的其他個案<br>• 必須定義相似性的測量，以找到適當的替代個案 | • 其他的個案可以取得<br>• 能夠確認適當的替代個案 |
| 熱卡 / 冷卡插補 | • 從最相似的個案或最佳的已知數值取得實際數值來替代遺漏資料 | • 必須定適合的相似個案或適當的外部數值 | • 確定替代的數值是已知的，或在相似性的基礎上，透過遺漏資料的處理找出適當的變數 |
| 隨機性遺漏資料處理的插補 | | | |
| 模型基礎法 | • 能處理非隨機和隨機的遺漏資料過程<br>• 是有最小偏差之數值的原始分布的最佳代表 | • 研究者才能詳細說明的複雜模型<br>• 需要專業的軟體<br>• 一般不是可以直接由軟體程式中取得（SPSS 的 EM 方法除外） | • 可以解決非隨機遺漏資料過程的唯一方法<br>• 資料的遺漏程度為高度且需要最小偏差的方法，以確保可通則化程度 |

**表 8.1　遺漏資料的插補技術比較（續）**

| 插補方法 | 優點 | 缺點 | 最佳使用時機 |
|---|---|---|---|
| 利用計算的替代值插補 | | | |
| 平均值替代 | ・易於了解及執行<br>・提供所有的個案有完整的資料 | ・減少分布的變異<br>・扭曲資料的分布<br>・削弱已觀察到的相關 | ・資料的遺漏程度相對較低<br>・變數之間有較強的關係 |
| 迴歸插補 | ・利用變數之間的真實關係<br>・以觀察個案在其他變數上所得到的數值為基礎計算替代數值<br>・每一個有遺漏資料的變數可以使用一組獨特的預測變數 | ・強化既有的關係和減少可通則化程度<br>・變數之間必須有充分的關係才能產生有效的預測數值<br>・除非將誤差項納入替代數值，否則會低估變異性<br>・替代數值可能「超出合理範圍」 | ・資料的遺漏程度為中度或高度<br>・變數間的關係必須充分確立，才不致於影響到可通則化程度<br>・軟體的可取得性 |

資料來源：Hairs et al. (2010, 55)

# 肆、資料來源、工具與研究設計

本章使用的資料來源為「臺灣選舉與民主化」(TEDS) 面訪資料。TEDS 是由行政院國科會支持的大型民意調查研究計畫，目標是有效運用有限資源記錄臺灣在民主化的過程中不同階段選舉前後民意的脈動與民眾政治偏好與行為。自 2001 年至 2012 年，TEDS 計畫收集了全國性與地方性選舉選後面訪調查實證資料，調查的議題包括選民投票行為、傳播接觸、社會動員、政治信任、政治功效感、政治參與、政策議題、施政滿意度、政黨認同、及個人基本資料等。[8]本研究所稱的「遺漏值」，指的是在

[8] TEDS研究調查為行政院國科會多年期、跨校的研究計畫，2004年計畫主持人為政治大學劉義周教授(NSC93-2420-H-004-005-SSS)、2008年計畫主持人為國立臺灣大學朱雲漢教授(NSC96-2420-H-002-025)，2012年為國立政治大學黃紀教授(NSC100-

原始資料檔中受訪者回答「不知道」、「拒答」、「很難說」、「無意見」等選項，亦稱作「無效值」，經過變數清理的過程後，在資料檔中該題的數值為空白、成為待插補的空格。

由於 TEDS 大型面訪案在全國性選舉的題目設計上較具有連貫性，因此我們選擇總統大選的資料作為應用多重插補方法的對象，使用 2004 年 (TEDS2004P, N = 1,823)、2008 年 (TEDS2008P, N = 1,905) 和 2012 年 (TEDS2012, N = 1,826) 面訪資料進行多重插補。值得說明的是，這三個資料檔的收集時間皆在總統選舉過後：2004 年的大選日為 3 月 20 日，TEDS2004P 面訪執行期間為 2004 年的 7 月中到 9 月；2008 年大選日為 3 月 22 日，TEDS2008P 執行期間為 2008 年的 6 月到 8 月；2012 年大選日為 1 月 14 日，TEDS2012 執行期間為 2012 年的 1 月中到 3 月初。另一方面，這些資料檔皆經過性別、年齡、教育程度、地區等人口學變數反覆加權 (raking) 的處理，以期與母體的在這些變數上接近。但是，由於樣本無法就政黨傾向與母體進行比對（因為沒有母體資料），因此，在人口學上與母體相近的樣本，不一定在其他變數面向上具有代表性。舉例而言，就「投票率」來看，由三筆面訪資料檔所計算出的投票率，比真實的投票率「底牌」高出一成至一成五左右。[9]

在插補處理流程上，首先，本研究利用「在這一次總統大選中，... 請問您有沒有去投票？」這個題目篩選留下「有去投票」的受訪者，就他們「投給誰」進行觀察，得到未作多重插補的投票分配。接著，本研究針對「投給誰」這一題未回答的受訪者予以多重插補的處理，再就完整的、已插補後的資料檔重新計算投票分配。由於使用商用軟體來處理多重插補的所費不貲，[10] 因此，本研究選用開放且免費的統計語言軟體 R，結合由哈

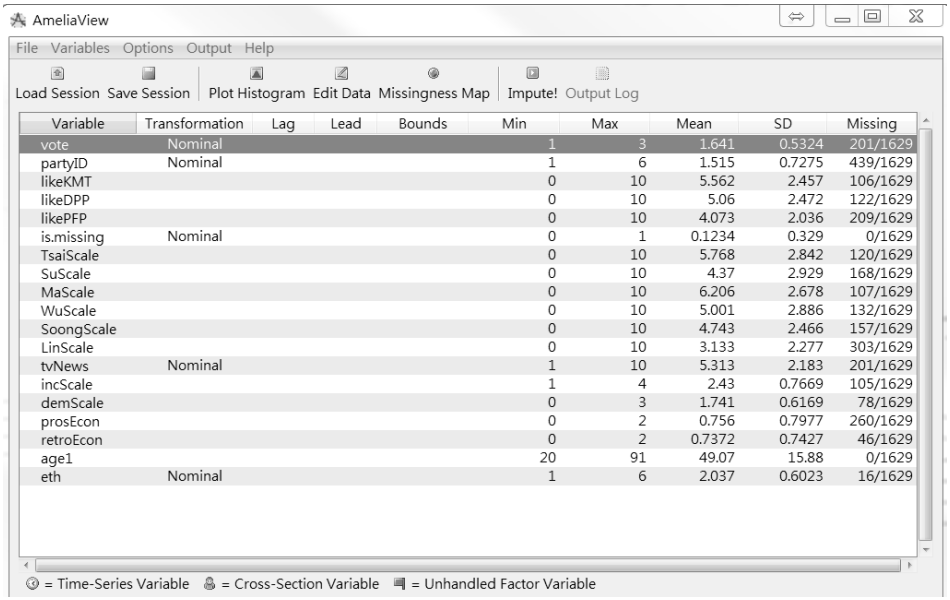
2420-H-002-030)；詳細資料請參閱TEDS全球資訊網：http://www.tedsnet.org。作者感謝上述機構及人員提供的資料，資料分析結果由作者自行負責。

9 TEDS2004樣本中的的總統大選投票率為91.23%，較實際投票率為高(74.38%)。TEDS2008樣本中的的總統大選投票率為88.37%，較實際投票率為高(80.28%)。TEDS2012樣本中的的總統大選投票率為89.36%，亦較實際投票率為高(74.38%)。

10 SPSS19.0版的多重插補模組價格單價就超過三萬元台幣。

佛大學政治學者 Gary King 等人所研發、貢獻的免費套件 Amelia II 來完成這個針對投票偏好遺漏值插補的工作。

*Amelia II* 使用 EMis(Expectation Maximization with importance re-sampling) 的演算法，使用了馬可夫鏈蒙地卡羅法 (Markov chain Monte Carlo，MCMC) 進行插補運算，較其他演算法如 chain equations 速度要快 (Honaker, King, and Blackwell 2009, 2011)。除了演算速度之外，這套軟體也較其他套件擁有容易操作的圖形化介面 (graphical user interface, GUI)，減少使用者在設定上的不便，也讓多重插補在技術上更為親民。例如，過去使用者往往必須忍受繁鎖的語法來設定每個輔助變數的類型為連續、順序或類別型變數（尤其重要的是要標示出類別型變數）。在 *Amelia II* 軟體中，使用者只要使用滑鼠及右鍵點選即可完成設定，且會提供各個變數的描述性資訊，包括最小值、最大值、平均數、標準差及遺漏資料數量等（操作畫面如圖 8.1 所示）。




| Variable | Transformation | Lag | Lead | Bounds | Min | Max | Mean | SD | Missing |
|---|---|---|---|---|---|---|---|---|---|
| vote | Nominal | | | | 1 | 3 | 1.641 | 0.5324 | 201/1629 |
| partyID | Nominal | | | | 1 | 6 | 1.515 | 0.7275 | 439/1629 |
| likeKMT | | | | | 0 | 10 | 5.562 | 2.457 | 106/1629 |
| likeDPP | | | | | 0 | 10 | 5.06 | 2.472 | 122/1629 |
| likePFP | | | | | 0 | 10 | 4.073 | 2.036 | 209/1629 |
| is.missing | Nominal | | | | 0 | 1 | 0.1234 | 0.329 | 0/1629 |
| TsaiScale | | | | | 0 | 10 | 5.768 | 2.842 | 120/1629 |
| SuScale | | | | | 0 | 10 | 4.37 | 2.929 | 168/1629 |
| MaScale | | | | | 0 | 10 | 6.206 | 2.678 | 107/1629 |
| WuScale | | | | | 0 | 10 | 5.001 | 2.886 | 132/1629 |
| SoongScale | | | | | 0 | 10 | 4.743 | 2.466 | 157/1629 |
| LinScale | | | | | 0 | 10 | 3.133 | 2.277 | 303/1629 |
| tvNews | Nominal | | | | 1 | 10 | 5.313 | 2.183 | 201/1629 |
| incScale | | | | | 1 | 4 | 2.43 | 0.7669 | 105/1629 |
| demScale | | | | | 0 | 3 | 1.741 | 0.6169 | 78/1629 |
| prosEcon | | | | | 0 | 2 | 0.756 | 0.7977 | 260/1629 |
| retroEcon | | | | | 0 | 2 | 0.7372 | 0.7427 | 46/1629 |
| age1 | | | | | 20 | 91 | 49.07 | 15.88 | 0/1629 |
| eth | Nominal | | | | 1 | 6 | 2.037 | 0.6023 | 16/1629 |

**圖 8.1　Amelia II 操作介面（以 TEDS2012 為例）**

使用者在完成變數類型的設定後，便可自動的將指定的類別型變數視為虛擬變數完成運算，產生使用者所預定要得到的多重插補後的資料檔數量。在本研究中每個 TEDS 資料檔皆各自產生 10 個完整的資料檔（第一階段），之後我們使用 R 程式語法，為每個資料檔計算每個政黨的得票率（第二階段），最後分別將 2004、2008、2012 三個年度各 10 個插補資料檔的政黨得票率取平均數，並計算標準差（第三階段）。[11]

第一階段（產生多組無遺漏值的資料檔）最重要的準備工作為選擇適當的參考變數，亦即與「投票偏好」具有相關性的解釋變數。為求比較上有共同的基礎，我們儘量選擇每個資料檔都有的、與政黨傾向或投票偏好相關的變數，包括了政黨傾向、喜好國民黨的程度、喜好民進黨的程度、喜好各組總統候選人的程度、喜好各組副總統候選人的程度、對過去臺灣經濟的觀感、對未來臺灣經濟的觀感、對當時政府執政團隊的觀感、對臺灣民主的滿意度、年齡與（受訪者父親的）省籍。[12] 表 8.1 至表 8.3 分別呈現了 TEDS2004P、TEDS2008P、以及 TEDS2012 資料檔所使用的變數名稱、代碼以及遺漏的程度。我們可以發現，政黨傾向與投票偏好這兩題不回答的比例非常高。政黨傾向題的無效值比例在 2004 年的資料檔中占了 67%，TEDS2008P 占了 33%，在 TEDS2012 占了 27%；投票偏好題的無效值比例在 2004 年的資料檔中占了 17%，TEDS2008P 占了 14%，在 TEDS2012 占了 12%。

---

[11] 另一個由Gary King等人開發的R套件Zelig可以將Amelia所產生出的多個資料檔進行分析及多個迴歸係數合併、計算平均數及標準差的工作(Imai, King, & Lau, 2004)。由於本研究只在計算點估計的平均值，用一般R語法就可達成，所以未使用此套件。

[12] 在2012年的資料檔中加入了喜好親民黨的程度，而「最常看的電視新聞台」在2004年TEDS中未包括，故只用在2008及2012的資料檔。

**表 8.1　投票選擇的多重插補變數列表** (TEDS2004P)

| 變數 | 變數全稱 | 變數代碼 | 遺漏數 |
|---|---|---|---|
| Vote | 2004 年總統大選投票選擇 | VH1B | 314 |
| partyID | 政黨傾向 | VH7A | 1218 |
| likeKMT | 喜好國民黨的程度 | VP2A | 147 |
| likeDPP | 喜好民進黨的程度 | VP2B | 144 |
| likePFP | 喜好親民黨的程度 | VP2C | 177 |
| ChengScale | 喜好陳水扁的程度 | VK11A | 108 |
| RuiScale | 喜好呂秀蓮的程度 | VK11B | 132 |
| LienScale | 喜好連戰的程度 | VK11C | 160 |
| SoonScale | 喜好宋楚瑜的程度 | VK11D | 158 |
| retroEcon | 對過去臺灣經濟的觀感 | VD01 | 204 |
| prosEcon | 對未來臺灣經濟的觀感 | VD02 | 373 |
| incScale | 對扁政府執政團隊的觀感 | VE01 | 164 |
| demScore | 對臺灣民主的滿意度 | VF01 | 204 |
| age1 | 年齡 | VS01 | 0 |
| Eth | 父親的省籍 | VS02 | 43 |

資料來源：黃秀端（2004）。
說明：N = 1,656（樣本中實際有去投票的人數）

**表 8.2 投票選擇的多重插補變數列表** (TEDS2008P)

| 變數 | 變數全稱 | 變數代碼 | 遺漏數 |
|---|---|---|---|
| Vote | 2008 年總統大選投票選擇 | H1a | 240 |
| partyID | 政黨傾向 | N1b | 550 |
| likeKMT | 喜好國民黨的程度 | N2 | 134 |
| likeDPP | 喜好民進黨的程度 | N2a | 146 |
| HsiehScale | 喜好謝長廷的程度 | J6a | 145 |
| SuScale | 喜好蘇貞昌的程度 | J6b | 156 |
| MaScale | 喜好馬英九的程度 | J6c | 147 |
| Siew Scale | 喜好蕭萬長的程度 | J6d | 168 |
| tvNews | 最常看的電視新聞台 | A3 | 207 |
| retroEcon | 對過去臺灣經濟的觀感 | E1 | 60 |
| prosEcon | 對未來臺灣經濟的觀感 | E2 | 319 |
| incScale | 對陳水扁執政團隊的觀感 | C1 | 145 |
| demScore | 對臺灣民主的滿意度 | F2 | 152 |
| age1 | 年齡 | S1 | 0 |
| Eth | 父親的省籍 | S2 | 15 |

資料來源：游清鑫（2008）。

說明：1. N = 1,680（樣本中實際有去投票的人數）；

2. 就 tvNews 這個變數，我們只挑前十個最常被接觸的電視新聞台，其他的就列為第十一類，完全沒有接觸電視新聞台的則編碼 0。

**表 8.3　投票選擇的多重插補變數列表** (TEDS2012)

| 變數 | 變數全稱 | 變數代碼 | 遺漏數 |
|---|---|---|---|
| Vote | 2012 年總統大選投票選擇 | H01a | 201 |
| partyID | 政黨傾向 | Q01b | 439 |
| likeKMT | 喜好國民黨的程度 | Q02 | 106 |
| likeDPP | 喜好民進黨的程度 | Q02a | 122 |
| likePFP | 喜好親民黨的程度 | Q02b | 209 |
| TsaiScale | 喜好蔡英文的程度 | J02a | 120 |
| SuScale | 喜好蘇嘉全的程度 | J02b | 168 |
| MaScale | 喜好馬英九的程度 | J02c | 107 |
| WuScale | 喜好吳敦義的程度 | J02d | 132 |
| SoonScale | 喜好宋楚瑜的程度 | J02e | 157 |
| LinScale | 喜好林瑞雄的程度 | J02f | 303 |
| tvNews | 最常看的電視新聞台 | A03 | 201 |
| retroEcon | 對過去臺灣經濟的觀感 | E01 | 46 |
| prosEcon | 對未來臺灣經濟的觀感 | E02 | 260 |
| incScale | 對馬政府執政團隊的觀感 | C01 | 105 |
| demScore | 對臺灣民主的滿意度 | F05 | 78 |
| age1 | 年齡 | S01 | 0 |
| Eth | 父親的省籍 | S02 | 16 |

資料來源：朱雲漢（2012）。

說明：1. N = 1,629（樣本中實際有去投票的人數）；

2. 就 tvNews 這個變數，我們只挑前十個最常被接觸的電視新聞台，其他的就列為第十一類，完全沒有接觸電視新聞台的則編碼為 0。

# 伍、結果與分析

以下就這三次總統大選選後的插補結果作出比較與報告。首先，我們先觀察遺漏值的插補情形。表 8.4 呈現的是「投票選擇」未答者的投票

偏好經插補後所呈現的分配情形，總體來看，插補並未造成某個政黨的支持者比例爆增的現象。在 TEDS2004P 的 314 名未告知其投票選擇的受訪者中，約有 48%「可能」投給了國民黨的候選人，約有 49%「可能」投給了民進黨的候選人。而 TEDS2008P 中，未回答投票選擇的受訪者中，「可能」為民進黨支持者的比例 (53%) 較「可能」為國民黨支持者的比例 (47%) 為高出一些。這情形也在 TEDS2012 中重現：在 201 位拒絕告知投票給誰的受訪者中，47%「可能」為民進黨的支持者，43%「可能」為國民黨支持者，10%「可能」為親民黨（宋楚瑜、林瑞雄）的支持者。由此可知，在投票偏好上未表態的受訪者「可能」是民進黨支持者的比例略高於「可能」是國民黨支持者的比例，亦即有較高比例之民進黨支持者在調查中隱藏了投票傾向。

**表 8.4 「投票選擇」未答者的投票偏好經過多重插補後所呈現的分配情形 (%)**

| 資料檔 | TEDS2004P | TEDS2008P | TEDS2012 |
|---|---|---|---|
| 國民黨 | 47.92<br>(0.01) | 47.33<br>(0.03) | 43.53<br>(0.05) |
| 民進黨 | 48.72<br>(0.01) | 52.67<br>(0.03) | 46.92<br>(0.05) |
| 親民黨 | - | | 9.55<br>(0.02) |
| 未答者總數 (N) | 314 | 240 | 201 |

資料來源：黃秀端（2004）、游清鑫（2008）、朱雲漢（2012）。
說明：括弧中數值為標準差

其次，本研究將以上這些原來不表態的民眾，其投票偏好經過多重插補法的計算「猜出」後，加回原始的資料檔與其他已經表示投票對象的民眾重新計算政黨得票率，如表 8.5 所整理。表 8.5 呈現了：(1) 未進行插補、純就樣本有效值所計算出來的各黨得票率、(2) 考量所有受訪者（包含偏好被「虛擬還原」的受訪者）的投票偏好計算出的得票率，與 (3) 當年度各組候選人實際得票率併列作初步的比較。圖 8.2 則是以柱狀圖的方式呈

現表 8.5 的內容。若就未進行多重插補的原始數據比對大選結果得票率來看，TEDS2004P 明顯低估了國民黨的得票率、稍高估了民進黨的得票率，2008 及 2012 的樣本則是稍低估了民進黨得票率而稍高估了國民黨的得票率。因此，原始的資料檔雖然在人口學變數上可以代表母體，但在政黨傾向上則尚未能準確反應得票的分配，當然，這或許也可能與調查時間點是在選舉結束後半年左右，可能有些受訪者因為社會期許考量而隱藏其真實之投票行為，傾向回答自己是投票給選舉中的勝選者有關。

不過，若就多重插補後的表現來看，本次研究總共進行了七組多重插補的計算，除了 TEDS2012 在計算國民黨得票率時標準差稍高 (0.63) 之外，其他六組使用 10 筆資料進行多重插補計算時的標準差都低於 0.01，表示每次進行投票偏好的插補運算其結果變異不大。由表 8.5 結合圖 8.2 來看，除了 2012 年親民黨的得票率被「調整過頭」（比實際得票率高了 0.75 百分點）之外，多重插補之後得到的政黨得票率都得到了校正。換言之，校正結果較插補前要更接近直接的開票結果。其中校正幅度最高前二者為 8.94%（TEDS2008P 國民黨）與 7.33%（TEDS2004P 國民黨）。整體來說，從多重插補結果與開票結果的比對來看，以 2004 年與 2008 年校正

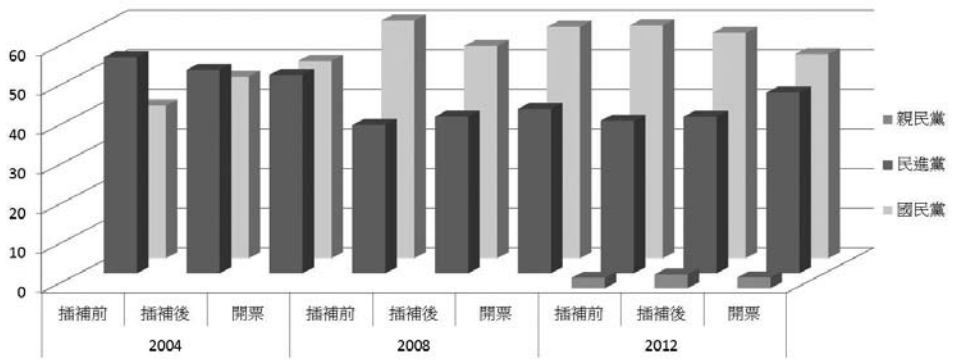
**表 8.5　使用多重插補法前後得票率估算結果比較 (%)**

| 年度 | 2004 | | | 2008 | | | 2012 | | |
|---|---|---|---|---|---|---|---|---|---|
| 政黨 | 插補前 | 插補後 | 開票 | 插補前 | 插補後 | 開票 | 插補前 | 插補後 | 開票 |
| 國民黨 | 45.41 | 45.61 (0.004) | 49.89 | 62.57 | 53.63 (0.004) | 58.45 | 58.82 | 56.94 (0.006) | 51.60 |
| 民進黨 | 54.59 | 51.60 (0.004) | 50.11 | 37.43 | 39.61 (0.004) | 41.55 | 38.52 | 39.55 (0.004) | 45.63 |
| 親民黨 | - | - | - | - | - | | 2.66 | 3.51 (0.003) | 2.76 |

資料來源：黃秀端（2004）、游清鑫（2008）、朱雲漢（2012）。

說明 1. 括弧中數值為標準差。

2. 因為多重插補是針對各政黨的得票比例分別計算（是由 10 組資料取得的平均數，所以會有標準差），也因此針對各個政黨插補之後的比例加總不一定會剛好是 100%。




資料來源：黃秀端（2004）、游清鑫（2008）、朱雲漢（2012）。

**圖 8.2　使用多重插補法前後得票率估算結果比較圖**

後所得的結果最好，2012 年的結果即使經過校正，國民黨得票率仍被明顯高估、民進黨得票仍被明顯低估，有無可能是因為三組以上之候選人，插補的效果較不理想，有待後續更多研究的檢證。總而言之，相較於什麼都不做，多重插補在點估計上的效果是存在的。

## 陸、結論與討論

本研究比較使用多重插補與不插補（單純使用加權過後的百分比）的結果，得到兩個面向的結論：(1) 使用多重插補法用於描述統計上有助於增加點估計的精準度；(2) 未來學界及實務界可以嘗試採用多重插補法於變數的描述。這兩個結論是建立在分析 TEDS 最近三次總統大選選後面訪資料的結果上，多重插補結果並未出現太過誇張的校調結果，且與大選結果的「底牌」來對照，用多重插補法於描述統計的結果並沒有比較不好。換言之，若有使用樣本來描述母體特徵的需求、且在沒有相對應的母體參數可供加權時，使用多重插補法可以彌補因遺漏值高而可能出現的估計上的偏誤。

需要注意的是，多重插補法需要研究者自行確認所採取的參考變數（如本研究中所使用的對候選人的評價）與標的變數（即本研究中的投票偏好「投給誰」）之間具有邏輯上或理論上的關聯。雖然變數愈多愈有助於正確地插補數值，但隨意挑選變數將無助於校調的準確性。因此，本方法比較適合應用在已經有理論及實證研究成果可供變數設計（或挑選）參考的研究領域上，如選舉前觀察投票偏好的分布和非選舉期間觀察政黨傾向的分布變化等，都是未來可以應用的方向。即使如此，變數的選取仍十分主觀，不同的變數組合可能產生些微不同的結果。因此，未來使用這個方法的研究者必須揭露所挑選的變數。

本研究之後還有許多值得繼續探討的議題。首先是將樣本校正的概念應用到「選舉前」的資料上。本研究選擇以「選舉後」的面訪案來作應用，是由於 TEDS 面訪案具有相當嚴謹的抽樣及執行過程，且是面對面訪問，因此樣本代表性較一般電訪要來得高。只是，用選舉後的調查資料來「回顧」政黨得票率不具實際的預測效益，而只具有方法上的效益。目前已有使用選前的面訪案進行應用的例子 (Liu 2010)，我們期望看到未來有更多應用選前電訪資料，應用這個方法進行預測之後的結果分享。其次是多重插補方法應用到其他類型或層級的選舉上，檢證其校正效果。由於本研究所使用的是總統大選資料，未來是否能應用到地方選舉，以及相關的變數是否需要重新考量等，都需要進一步研究討論。

第三是這個方法的外部有效性檢測。我們使用 TEDS 資料的原因，是預設了它的樣本代表性，也就是相信樣本來自 100% 範圍的母體，才拿插補處理過後的樣本跟大選結果的「底牌」來對照。這樣做是較方便的作法，而非真正的插補的有效性的檢證作法。未來在驗證插補運算結果是否合理的研究中，我們期望進行外部有效性檢測，也就是以實驗的方式，在受訪者願意被再接觸的前提下，將插補所「猜」出來的結果與這些受訪者真實的答案作比對，如此才能真正確定這個方法的實用性，也有助於確認怎麼樣的參考變數的組合最具有預測力。

第四是遺漏假設的檢測。目前使用多重插補的研究者都必須預設受訪者不回答的情況是「隨機性遺漏值」(MAR)，且所找出來的參考變數就是

導致非隨機性遺漏的原因。然而，我們知道遺漏值型態還包括了「完全隨機遺漏值」(MCAR) 與「非隨機遺漏值」(NMAR)，只是目前並未有方法來診斷這個針對資料檔有關 MAR 的預設是否正確，這是個學界尚未突破的地方，也是未來應用多重插補方法的研究者必須留意之處。

第五是各種插補工具的比較。本章所介紹插補法中，雖然多重插補法是比較成熟、考量面向較完整的作法，但是多重插補法中仍有多種運算方式、套件工具等有待未來研究者進行比較與嘗試。例如在 R 及 Stata 軟體上都可以使用的 MICE 套件也具有相當成熟的插補運算式，也相當值得研究者進行使用、比較與分享。本章推薦的 Amelia II 軟體其優點除了免費及操作容易外，運算速度也較 MICE 來得快。

最後是插補效果的檢討。本章所使用的資料是有去投票的受訪者的投票偏好，但結果並未能更接近大選的結果。原因除了上述的限制之外，也可能是資料本身的在政黨偏好上的偏誤。以 TEDS2012 來看，我們發現樣本偏「藍」，即使經過插補後的資料國民黨的支持者仍然被高估。在沒有真實母體參數可供對照的情況下，我們推測有可能是抽樣造成的結果。

總括來說，多重插補法是處理項目無反應的資料時十分有用的工具。雖然它的有效性尚未得到全面的檢視，然本章使用三筆國內調查中資料品質最高的 TEDS 面訪資料所得到的結果肯定了這個方法用於校準點估計上的潛力。研究者未來在應用上必須考量文獻中已經點出的多重插補的限制，並同步注意抽樣代表性及樣本涵蓋率的問題，才能將樣本的預測能力及描述母體的能力充分發揮出來。

# 參考文獻

## I. 中文部分

伊慶春、蘇碩斌，1995，〈無作答之分析〉，載於《社會調查與分析》，章英華、傅仰止、瞿海源主編，台北：中央研究民族學研究所。

朱雲漢，2008，《2005 年至 2008 年『選舉與民主化調查』四年期研究規劃 (III)：2008 年立法委員選舉面訪案》，計畫編號：NSC 96-2420-H-002-025，台北：行政院國家科學委員會補助專題研究計畫成果報告。

------，2012，《2009 年至 2012 年『選舉與民主化調查』三年期研究規劃 (3/3)：2012 年總統與立法委員選舉面訪案》，計畫編號：NSC 100-2420-H-002-030，台北：行政院國家科學委員會補助專題研究計畫成果報告。

杜素豪，2004，〈投票意向問題不同類型項目無反應之分析：以 2000 年總統大選為例〉，《選舉研究》，11(2): 111-131。

邱皓政，2003，《結構方程模式：LISREL 的理論、技術與應用》，（初版），台北：雙葉。

許玉雪、林建弘，2008，〈插補法在不同缺失機制下之比較分析〉，《統計與資訊評論》，10: 19-39。

陳信木、林佳瑩，1997，〈調查資料之遺漏值的處置——以熱卡插補法為例〉，《調查研究》，3: 75-106。

游清鑫，2008，《2005 年至 2008 年『選舉與民主化調查』四年期研究規劃 (IV)：2008 年總統選舉面訪案》，計畫編號：NSC 96-2420-H-004-017，台北：行政院國家科學委員會補助專題研究計畫成果報告。

黃秀端，2004，《2002 年至 2004 年『選舉與民主化調查』三年期研究規劃 (III)：民國九十三年總統大選民調案》，計畫編號：NSC 92-2420-H-031-004，台北：行政院國家科學委員會補助專題研究計畫成果報告。

黃紀，2012，《2009 年至 2012 年『選舉與民主化調查』三年期研究規劃 (3/3)：2012 年總統與立法委員選舉電訪調查》，計畫編號：NSC 100-2420-H-002-030，台北：行政院國家科學委員會補助專題研究計畫成果報告。

鄒慧英、江培銘，2012，〈插補法在檢測試題差異功能的效果〉，《測驗學刊》，59(1): 1-32。

廖培珊、江振東、林定香、李隆安、翁宏明、左宗光，2011，〈葛特曼量表之拒答

處理：簡易、多重與最鄰近插補法的比較〉，《臺灣社會學刊》，47: 143-178。

劉義周，2004，《2002 年至 2004 年『選舉與民主化調查』三年期研究規劃 (IV)：民國九十三年立法委員選舉大型面訪案》，計畫編號：NSC 93-2420-H-004-005-SSS，台北：行政院國家科學委員會補助專題研究計畫成果報告。

## II. 英文部分

Berinsky, Adam J. 2008 “Survey non-response.” In *The Sage Handbook of Public Opinion Research* ,eds. Donsbach Wolfgang and Michael W. Traugott. Los Angeles: Sage.

Chen, Yinzhong, and Jun Shao. 1999. “Inference with Survey Data Imputed by Hot Deck When Imputed Values are Nonidentifiable.” *Statistica Sinica* 9: 361-384.

Dempster, A. P., Laird, N. M., and Rubin Donald B. 1977 “Maximum Likelihood From Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society,* 39(1): 1-38.

Dillman, Don A., John L. Eltinge, Robert M. Grove, and Roderick J. A. Lilttle. 2002 “Survey Nonresponse in Design, Data Collection, and Analysis.” In *Survey Nonresponse*, eds. Robert M. Grove, Don A. Dillman, John L. Eltinge, Roderick J. A. Little, New York: John Wiley & Sons, Inc.

Filion, F. L. 1976. “Estimating Bias Due to Nonresponse in Mail Surveys.” *The Public Opinion Quarterly* 39(4): 482-492.

Fix, Evelyn, and J. L. Hodges. 1951 “Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties.” Project 21-49-004, Report NO.4, US Air Force School of Aviation Medicine, Randolph Field.

Hair, Jr. Joseph F., William. C. Black, Barry J. Babin, and Rolph E. Anderson. 2010, *Multivariate Data Analysis: A Global Perspective* 7th ed. New Jersey: Prentice Hall.

Honaker, James, King Gary, and Blackwell Matthew. 2009. “Amelia Software Web Site.” http:// gking.harvard.edu/Amelia” (accessed April 24, 2009).

Honaker, James, King Gary, and Blackwell Matthew. 2011. “Amelia II: A Program for Missing Data.” *Journal of Statistical Software 45*(7): 1-47.

Imai, Kosuke, Gary King, and Olivia Lau. 2004. “Zelig: Everyone’s Statistical Software.” http://GKing.Harvard.Edu/zelig

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley and Sons, Inc.

Little, Roderick J. A., and Rubin, Donald B. 2002. *Statistical Analysis with Missing Data* 2nd ed. New York: John Wiley & Sons.

Liu, Frank C.S. 2010. "Reconstruct Partisan Support Distribution with Multiply Imputed Survey Dada: A Case Study of Taiwan's 2008 Presidential Election." *Survey Research* 24: 135–162.

Lohr, Sharon L. 2010. *Sampling: Design And Analysis* 2nd ed. MA: Duxbury Press.

Marker, David A., David R. Judkins, and Marianne Winglee. 2002. "Large-Scale Imputation for Complex Surveys." In *Survey Nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge, Roderick J. A. Little. New York: John Wiley & Sons.

Mason, Robert, Virginia Lesser, and Michael W. Traugott. 2002. "Effects of Item Nonresponse on Nonresponse Error and Inference." In *Survey Nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge, Roderick J. A. Little. New York: John Wiley & Sons.

Peugh, James L., and Craig K.Enders. 2004. "Missing data in educational research: A review of reporting practices and suggestions for improvement." *Review of Educational Research* 74(4): 525-556.

Rubin, Donald B., 1987. *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.

Rubin, Donald B., and Elaine Zanutto. 2002. "Using Matched Substitutes to Adjust for Nonignorable Nonresponse through Multiple Imputations." In *Survey Nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge, Roderick J. A. Little. New York: John Wiley & Sons.

Tanner, Martin A., and Wing Hung Wong. 1987 "The Calculation of Posterior Distributions by Data Augmentation (with Discussion)." *Journal of the American Statistical Association* 82: 528-550.