研究論文

調查資料之遺漏值的處置——以熱卡插補法爲例

陳信木*　林佳瑩*

摘要

遺漏值(missing values)之出現,似乎是處理社會學調查資料時,所面臨之不可避免難題。事實上,長久以來,社會學家早已體認此一事實,然而,多年來,社會學家進行實證研究分析之際,面對遺漏值的議題,除了少數嫻熟統計理論與電腦程式的學者之外,往往採取『棄之不理』的策略。固然大家深知此種處置方式即將引發諸多困境,諸如選擇偏誤(selection bias)、甚或樣本規模嚴重流失等,但是,既存相關的統計文獻,普遍晦澀艱深,同時,目前盛行的電腦統計軟體並未提供簡易套裝程式直接處理遺漏值問題。導致結果,社會學家雖然深刻體認必須處置遺漏值問題,可是,現實上却又無從追隨,只有忍痛割愛。

這篇論文的主旨,就是一般性地介紹討論遺漏值處置議題。首先探討社會學調查資料中所出現的遺漏值課題,尤其指出遺漏值可能引發的一些困境。其次,我們介紹若干遺漏值的處置方式(treatments)或是策略;然後,本文以實際的調查資料爲例,應用所謂熱卡插補方法,處置調查資料中的遺漏值。

* 二位作者爲政治大學社會學系副教授。

出現「遺漏值」(missing values)或是「不完整資料」(incomplete data) 的問題，幾乎是所有調查資料 (survey data) 不可避免的難題。目前，社會學的經驗研究，相當仰賴調查方法和調查資料，尤其深受此一困擾。然而，即使社會學家已經體認此一事實，長久以來，或許囿限於相關統計理論文獻晦澀艱深，加以缺乏實際處理遺漏值資料的經驗，[1] 導致結果，面對遺漏值的議題，往往採取「棄之不理」的策略，或是抱持「視而不見」的鴕鳥式作法，不幸地，卻將引發諸多困境，因此，社會學家在資料分析過程，實在不能對這個問題忍痛割愛。本篇論文的主旨，就是一般性介紹討論社會學調查資料的遺漏值處置課題，並且，以常用的「熱卡」插補法展示說明，希望藉此拋磚引玉，鼓勵社會學研究者勇於面對此一困難。

不過，在此，我們首先必須指出本文研究的性質以及限制。本篇論文的目的，主要是爲重新引發社會學研究者體認調查資料中的遺漏值問題，特別加以重視之，是以，本文論述將不以相關的統計理論爲主。此外，本文討論的方向，係以社會學調查資料爲主，對於其他研究設計方式所獲得的資料，特別是實驗設計以及此類資料中的遺漏值議題，則不加探討。[2]

---

1 目前，社會科學界廣泛被採用於資料分析的統計套裝軟體 (例如 SAS 和 SPSS)，只有納入 pair-wise deletion 和 list-wise deletion 兩種遺漏值處置方法，可能也是導致社會學研究者忽略遺漏值處置重要性的原因之一。

2 經由實驗 (experiment) 設計所獲得的研究資料，當然，也不可避免地產生遺漏值問題。實驗設計資料的統計分析，傳統上是以變異數分析 (analysis of variance) 爲主軸，因此，此類遺漏值的處置策略，大多數也以這一系列的方法作爲重心。然而，典型的社會學研究，並不偏好這種「版本」(version) 的統計分析技術，是以，本文不擬加以討論。不過，許多的理論文獻，對於這一類型的遺漏值處置技術，多有深入淺出的討論介紹 (例如，Anderson, Basilevsky, and Hum, 1983; Armimger and

壹、社會學調查資料中的遺漏值課題

顯然地，社會學研究者早已體認分析資料中的遺漏值問題，可是，有意或是無意地，卻加以忽視——即使是謹嚴的研究者，往往只是在研究報告中輕描淡寫地告知讀者「the missing data were ignored」或是「only complete cases were used in the analysis」。[3]

一般來說，調查資料中所出現的遺漏值或是不完整資料，就是通稱的「無反應」問題，可以區分爲兩類型：單位無反應（unit nonresponse）以及項目無反應（item nonresponse）。單位無反應是指樣本中的部份觀察體之全部資訊遺失，項目無反應則是樣本中某些觀察體的部份（變項）資訊遺失。

造成單位無反應的原因，主要和調查方法（設計）與資料蒐集過程（例如訪問）有關，至於其所衍生的嚴重後果，特別是產生所謂樣本代表性（representativeness）問題。因此，有關調查資料中的單位無反應問題，普遍受到社會學研究者熱切關注，相關的研究論述眾多（例如，章英華、傅仰止、瞿海源，1995；洪永泰，1986, 1989, 1995），也提出許多預防及治療的處置方法（例如，楊文山、蔡瑤玲，1995；Bailey, Chapman, and Kasprzyk, 1986; 詳見 Lessler and Kals-

---

Sobel, 1990; Little and Rubin, 1987, Chapter 2; McArdle, 1994; Mendoza, 1993; Mislevy, 1993; Muthn, Kaplan, and Hollis, 1987; Rovine and Delaney, 1990）。

3 Afifi and Clark（1990:223-25）的著名參考工具書《Computer-Aided Multivariate Analysis》就是指導使用者利用統計軟體既有的 pair-wise 或 list-wise deletion 方法，簡單地處理遺漏值問題，當然只能告訴讀者：研究者的確「知道」分析資料中存在遺漏值的問題。

beek, 1992:163-207 的討論)。本文將不以此類的單位無反應問題作爲對象，下文探討重心也將對之略而不談。

至於項目無反應的問題，姑且不論產生的原因，由於這個現象非常普遍出現於調查資料中，而且，也是資料分析過程直接面對的困擾，因此一般討論遺漏值課題之時，大都以它爲對象。通常，標準的統計分析程序係以所謂「矩形的資料結構」(rectangular data)[4] 作爲對象，如果資料當中出現項目無反應的案例，即將破壞矩形的資料結構，進一步導致一些統計的問題。

具體言之，部份觀察體出現項目無反應，致使研究資料損失，因而減低統計能力 (statistical power)。其次，項目無反應的案例，由於破壞完整的矩形資料結構，可能導致資料分析最後所估計的參數值 (parameter estimate) 變成偏誤 (biased)。以包含遺漏值的資料作爲分析對象時，不僅可能相當程度地導致模型估計的參數值偏誤，即使是單變項的描述統計值，也可能嚴重偏誤。

最後，社會學研究者在未來進行資料分析過程之時，所面臨的遺漏值課題必然更加困擾、嚴重。究其緣故，目前的社會學調查資料，大多數是橫斷面性質 (cross-sectional)，即使出現遺漏值問題，只是影響該橫斷面時空點上的樣本。但是，社會學研究日漸朝向時貫性研究 (longitudinal) 的方向前進，而且，大量的長期性調查資料出土問世，更加指出社會學家在未來必須處理遺漏值課題的迫切性。簡單來說，時貫性資料，乃是由系列的橫斷面資料串連而成，每一波 (wave) 橫斷面時空點上的遺漏值案例，將會導致此個案在全程的觀察期間損

4 所謂矩形的資料結構，典型地，就是以「觀察體×變項」(cases variables) 方式所安排而成的資料矩陣。

耗（attrition），最後致使樣本的全部資訊趨向「相乘、遞減」，嚴重惡化遺漏值問題。正就是因爲如此，社會學的長期性研究，特別重視遺漏值的課題（Chowdhury, 1991; Fay, 1989; Lepkowski, 1989; Little, 1992a; Little and Su, 1989; Marini, Olsen, and Rubin, 1980）；無怪乎進行長期性研究的社會學家，第一個任務，就是必須殫盡心力於處理遺漏值問題（例如 Hayward and Grady, 1990）。

無論如何，遺漏值的確是社會學家進行資料分析過程必須重視的課題，而且，任何調查資料不可避免地將出現遺漏值。那麼，如何處置分析資料中的遺漏值問題呢？本文下節論述，探討若干的遺漏值處置方法——對於這些方法的討論內容，將從社會學研究的實用角度出發，並不深入涉及統計理論內涵。

貳、若干的遺漏值處置方法

論及遺漏值的處置方法，人類亙古的智慧——「預防勝於治療」，永遠是唯一絕佳的策略。也就是說，處置遺漏值的最好、最有效方法，就是避免資料出現遺漏值的案例。很不幸地，無論如何避免，幾乎任何的調查資料，或多或少都會出現遺漏值。是故，遺漏值一旦出現，如何加以治療性處置，就成爲資料分析過程的重要任務。[5]

首先，各種遺漏值的處置措施之間差異，源於對產生遺漏值的機轉（mechanism）抱持不同的預設。有時候，遺漏值產生的機轉是在研究者掌控之下，則此機轉乃是「可忽略的」（ignorable）。例如，在

5 關於遺漏值的預防性處置措施，大多數的調查方法論之相關文獻以此爲重心，本文不另加討論之。

機率樣本抽取過程，未被納入樣本的個案，其資料必然是遺漏的，所以，導致非樣本中的個案資料遺漏之機轉，就是可以被忽略的。

然而，很多時候，產生遺漏值的機轉卻不是在研究者掌控之下，因此，也不可以忽略（nonignorable）。例如，「censoring」現象可能導致資料遺漏，就是一例。如果產生遺漏值的機轉不可以忽略，研究者勢必對其訂立若干預設。舉例來說，一組資料包含 K 個變項，那麼，某一個案的 X_1 變項數值遺漏的可能性，或許①和整組資料完全無關，或許②端視該個案的 X_1 數值而定，或許③取決於 X_2 的數值，或許④取決於 $X_2, X_3 \cdots\cdots X_k$ 的數值。

上述第一種現象，稱之爲「完全隨機遺漏」（missing completely at random; MCAR），亦即，某一個案的特定變項數值遺漏之可能性，獨立於整組資料，係爲完全隨機之下產生。第三種和第四種現象，則只是「隨機遺漏」（missing at random; MCR），意指某一個案之特定變項數值遺漏的可能性，取決於其他被觀測的變項之數值。至於第二種現象，也就是某一個案之特定變項數值遺漏的可能性，端視其數值而定（譬如，社會調查過程經常發現，高所得的受訪者傾向於拒絕回答其收入，因此，所得變項的數值與產生遺漏的可能性有關），則不是「隨機地遺漏」。

既有文獻中所討論有關遺漏值的處置措施，大皆預設遺漏值的出現、分佈呈現完全隨機（MCAR）的模式，本文以下討論也以此爲出發點。[6] Afifi and Elashoff (1966, 1967, 1969a, 1969b) 以及 Hartley

6 Cohen and Cohen （1983）、Gilley and Leone （1991）、Kim and Curry （1977）、Little and Rubin （1987）、Rubin （1976）等文獻，提出若干檢驗遺漏值模式是否隨機分佈的方法。

and Hocking (1971) 最早綜合回顧既有的遺漏值處置文獻，嗣後，許多的方法和策略不斷發展；在此，我們依循這些分類 (Chapman, 1976; Jinn and Sedransk, 1989; Kalton and Kasprzyk, 1982, 1986; Little, 1988, 1992b; Little and Rubin, 1989-90)，將遺漏值處置方法區分爲四個類別：

(1)完整的觀察體 (complete-case) 分析法
(2)加權 (weighting) 法
(3)插補 (imputation) 法
(4)模型建構法

其次，我們必須指出，遺漏值處置的工作通常在實際進入社會學模型建構之前進行，在此所謂的「多變項」，並非意指「多變項模型建構」(multivariate modeling)，而只是單純描述資料組包含若干的變項。所以，在此所謂的應變項 (dependent variable)，係指等待遺漏值處置的變項，而解釋變項則是輔助之用，兩者之間並未設定必然的社會學因果關連。

(1)完整的觀察體分析法

所謂完整的觀察體 (complete-case) 分析法，以最簡單的形容詞來說，就是「丟掉它」，亦即，一個觀察體，在統計計算過程，只要所涉及的任何一個變項出現遺漏值，則將之排除於分析之外，因此，統計分析時所使用的資料，都是完整的觀察體。這種方法，一般統計軟體稱之爲成批刪除 (list-wise deletion)，通常也是統計軟體的內設 (default) 方法。

以完整觀察體進行分析而處置遺漏值，的確是最簡單的方法，也被廣爲採用。但是，這個方法的致命之處，則是損失大量資料，最終

導致樣本偏誤。例如，Kim and Curry（1977）以五個變項的資料矩陣進行實驗，如果每個變項上皆有 10%的個案之觀察值遺失，且是隨機的，那麼，以成批刪除方法加以處置，結果只剩餘 41%的個案可以用於分析。爲了這個緣故，許多人使用「成對刪除」（pair-wise deletion）的方法處理遺漏值。[7]

所謂成對刪除，就是在進行統計計算時，一個個案即使其某一變項的觀察值遺失，仍將之保留，唯有當「必要時」才將它排除。相對於成批刪除方法，成對刪除可能避免較大量的資訊流失，不過，其代價則是產生不穩定的共變量矩陣（unstable covariance matrix），也就是共變量矩陣並非 positive-definite，因而應用共變量矩陣進行多變量分析時，發生統計的計算問題，甚至進行統計推論時也困難重重。[8]

無論如何，以完整的觀察體作爲分析對象的作法，固然產生資料流失、樣本偏誤、和若干的統計問題，這種方法由於「簡單」，且不涉複雜的統計理論，因而，始終也是社會學研究者進行資料分析時，處置遺漏值的「最愛」方法。

⑵加權法

個案的某一變項之觀察值遺失，將導致此個案在這個變項的分配

---

7 當然，也有研究者質疑，一旦遺漏值所佔全部資料的比例過高，那麼，究竟是否「合法地」應用各種處置的措施？抑或必須重新蒐集資料？對於這個質疑，我們無法提供解答，畢竟，此乃研究者必須評估的抉擇。本文的討論，係假設研究者別無選擇之下，必須援用「手上」現有的調查資料時，可能遭遇的有關遺漏值問題以及可以採用的處置方法。

8 Little and Rubin （1987:42-43）指出，在「完全隨機遺漏」條件下，成對刪除法仍可產生一致的共變數估計值。

上喪失代表（影響力）；是以，如果樣本中存在其他具有相同變項特徵的觀察體，那麼，藉由加權這些觀察體的代表性，可以補償此一特定觀察體資訊遺漏的損失。這就是加權法的基本精神。傳統上，加權法被廣爲應用於單位無反應的情境，甚至，許多的抽樣設計本身也充分運用加權設計而達到代表性和統計分析的目的（參見 Fuller, 1974; Lee, Forthofer, and Lorimor, 1986; Mandell, 1974）。舉例來說，電話訪問所得的樣本，通常是偏誤的，因爲男性的單位無反應率較高，那麼，研究者可以利用男性的反應率（response rate）加權樣本中的男性受訪者之代表性，以反映母體中的實際分配狀況。

當然，加權法也可以應用至項目無反應的情境，進一步處置遺漏值的問題。不過，目前社會學界較少採用加權法處置遺漏值，因爲：第一，加權數的設定，必須考慮反應率，很多時候，實際的反應率並不可知；第二，項目無反應的狀況中，每一特定項目（變項）之反應率可能不同，因而必須分別處理，如果分析模型涉及多個變項，益加使得問題更形複雜。因此，過於冗長、費時的加權計算過程，使得加權法並未普遍被採用於處置遺漏值。

⑶插補法

插補法的基本目的，就是一旦出現遺漏值時，則找尋一個數值替代之。由於找尋和替代的策略不同，目前實用的插補法眾多，並不限於單一方式——不過，所有插補法的共通目的，就是盡可能找尋一個和遺漏值相似的數值替代之。至於如何找尋，一般都是仰仗若干「輔助變項」(auxiliary variables) 所提供的資訊達到——不論是出現遺漏值的個案或未出現遺漏值的個案，如果兩者的輔助變項表現相近資訊（當然，輔助變項不可以出現遺漏值），那麼，我們就可以推論，兩

者在出現遺漏值的特定變項上，表現亦是接近，因而得以替代之。[9] 舉例來說，$X_1, X_2, \cdots X_k$ 就是輔助變項，個案 A 和 B 在這 k 個變項上的行爲表現相同，但是個案 A 在 X_{k+1} 變項的數值遺失，而個案 B 則可以觀察得到 X_{k+1} 變項的數值。藉由兩者在 $X_1, X_2, \cdots X_k$ 等輔助變項表現相同，我們推論，個案 A 和個案 B 在 X_{k+1} 變項的數值應該相同，因此，以個案 B 在 X_{k+1} 變項的數值替代個案 A 的流失資訊。

①平均數替代

平均數（mean）替代（插補）法的基本假設，就是在遺漏值的分配是完全隨機的前提之下，研究者相信，出現遺漏值的的若干觀察體，其在該變項的平均數，理論上相等於未出現遺漏值的觀察體之平均數，因此，我們以未出現遺漏值之觀察體的平均數替代所有遺漏值。平均數替代法可以說相當簡單運用，不過，也有一些缺陷。例如，可能扭曲 X 變項在樣本中的分配，因爲，所有出現遺漏值的觀察體，其 X 變項的數值只有一個，就是平均數（或是若干個條件性平均數）。這種事實，進一步也會降低、減少 X 變項的變異量，造成變異量低估的問題。

②迴歸法

迴歸法（regression method）應用下列方程式進行遺漏值插補：

$$Z = \beta_0 + \sum_{k=1}^{k} \beta_k X_k + \varepsilon$$

9 關於「輔助變項」的界定和選擇，通常與研究主題有關。有人認爲，輔助變項應該和被處理的遺漏值變項兩者有關連，如此才具有意義；相反地，也有人主張，這兩者之間不應過於密切關連，畢竟，因果關係的追尋，正就是研究的目的，所以，輔助變項最好不要和處置的變項之間存在直接關係。大體而言，目前，社會學領域中較爲普遍採用基本人口學和社會學變項（諸如年齡、性別、種族等）。

具體來說，研究者的關切重點是 Z 變項，但是，全部樣本中只有 m 個觀察體的 Z 變項未流失。現在，我們認爲 Z 變項與一組輔助變項 $(X_1, X_2, \cdots X_k)$之間存在線性關係，所以，首先就以 m 個觀察體計算上述的迴歸方程式，估計其迴歸參數值，然後，Z 變項出現遺漏值的觀察體，就以此迴歸方程式預測其 Z 變項數值。

一般來說，迴歸法較之成對和成批刪除方法更能保留資料（因爲，大部份的觀察體可以藉由迴歸插補而得能保留，毋需刪除之），而且，迴歸預測值容許離差值（deviation），因此，不致過於扭曲應變項的分配情形。不過，當研究者應用迴歸法進行插補遺漏值時，必須注意迴歸方程式所預測的數值是否合理，或是落入有效的範圍，也就是必須考量迴歸預測的外插問題（extrapolation）。當然，如果等待插補的變項係是類別性的(categorical)，上述的迴歸模式可以延伸使用一般化線性模型（generalized linear model），尤其是 log-linear 模型最常被應用。

除此之外，研究者可能認爲，在輔助變項上具有條件的若干個觀察體，其應變項（Z）存在差異，也就是彼此之間的迴歸預測值（Z'）雖然相同，其實際預測值則有差異；那麼，我們可以在上述的迴歸方程式中加入隨機的殘餘值（random residual），亦即容許 ε_i 不爲 0。至於指定 ε_i 數值的方法也有很多，例如，從 $N(0,\sigma_\varepsilon)$中隨機取出一個數值；[10] 如果研究者不能肯定迴歸模型的同質相等性（homoscedasticity），可以考慮輔助變項的不同條件，以相對應的 $N(0,\sigma_\varepsilon)$產生隨機殘餘值。另外，比較簡單的作法，則是在未出現遺漏值的觀察體中，

10 設定殘餘值的分配是常態的，平均數爲 0，標準差則是上述迴歸方程式的殘餘值之標準差，亦即 σ_ε。

隨機選出一個觀察體，以其迴歸殘餘值附加於被預測個案的迴歸預測值，作爲遺漏值的迴歸插補數值——其實，這個作法已經接近以下即將討論的熱卡插補法。

③熱卡插補法

熱卡插補（hot-deck imputation）法，可以說在目前的調查方法論中最受青睞，而著名的 CPS hot-deck 則已風行數十年。[11] 由於熱卡插補方法已經發展許多版本，我們綜合討論並簡單加以介紹（請參考 Ford, 1983; Rizvi, 1983; Sande, 1979, 1982, 1983）。

熱卡插補法[12] 的基本精神，就是按照輔助變項的不同條件，將未出現遺漏值的觀察體分類成爲若干的「插補空格」（imputation cell），然後，每一個出現遺漏值的觀察體，依據其輔助變項的條件，從相對應的「插補空格」中找尋一個觀察體，以其觀測所得的變項數值代替遺漏值。在此，輔助變項的指定，通常是基本的人口學變項或是社會學變項，諸如年齡、性別、種族、教育、社會經濟地位等；而且，若干個輔助變項所形成的眾多「插補空格」，必須是彼此周延

11 美國人口普查局的 Current Population Survey（CPS）和各種普查資料，廣泛應用熱卡插補方法處置遺漏值，經過數十年的努力，CPS hot-deck 可以說是社會科學調查資料之遺漏值處置措施中，最爲成熟的一種方法，當然也備受關注（David et al., 1986; Oh and Scheuren, 1980; Oh, Scheuren, and Nisselson, 1980; Welniak and Coder, 1980）。

12 從名詞的意義來看，熱卡插補和電腦應用密切相關。「卡」是一個電腦技術中的單位，指一綑打孔卡（punch cards）——每一個觀察體就是一個打孔卡，具有相同輔助變項的觀察體，形成獨一的「插補空格」，也就是組成一綑的打孔卡，而遺漏值插補過程就是從相對應的一綑打孔卡中取出一張，然後以其觀測所得之變項數值替代遺漏值。至於稱之爲「熱」的緣故，乃是因爲插補來源的這些打孔卡，和出現遺漏值的觀察體，都是隸屬相同的資料組——就是正在電腦中運算之「熱騰騰的」卡片。當然，相對而言，另外也有所謂的「冷卡」（cold-deck）插補，其替代數值的來源，則是來自其他資料的一綑打孔卡，並非手上正在處理的資料，故稱之爲「冷」。

(exhaustive)、互斥 (exclusive)、和同質的 (homogeneous)。

④其他方法

除了上述三類插補方法以外，另外也有一些曾經被採用、或建議的策略。例如，運用其他來源的資訊，進行「記錄配對」(record-matching) 的作法。舉例來說，美國人口普查局的 CPS 資料，曾經從十年人口普查、國稅局、監理處、或是社會安全等檔案中精準配對，進行遺漏值插補。當然，這種作法，除了必須擁有各種資料來源，實際運作過程的成本龐大，恐怕並非一般的社會調查研究者所能負擔。另外，也有人修正迴歸法和熱卡插補法，而提出所謂的「distance-function matching」方法 (Kalton and Kish, 1984; Sande, 1979, 1982)，以避免熱卡插補的困境。不過，由於相關的統計理論研究並不成熟發展，所以，較不被採用於調查資料分析。

⑷模型建構法

模型建構法，主要應用概度最大化 (maximum likelihood) 的統計理論，預設一個母體分佈的模型，然後，以觀測所得的樣本資料，在「概度最大化」的原則之下，估計參數值。近年來，應用 ML 理論或是 EM (expectation maximization) 理論，而進行遺漏值處置的研究，相當大量出現於統計研究領域 (例如，Dempster, Laird, and Rubin, 1977; Fuchs, 1982; Gelfand and Carlin, 1993)。這種方法具有許多優點，遠勝於其他的遺漏值處置方法。例如，只要模型是正確的，ML 的估計值是一致的 (consistent)、有效用的 (efficient)；其次，即使遺漏值的分佈並非完全隨機 (社會學的調查資料確實經常如此)，ML 的估計仍可能是一致的、有效用的；第三，由於這個方法應用 ML 理論，所以，能夠估計大樣本的標準誤，並且進行相關的檢定

和估計。

然而，事實上，以模型建構而處置遺漏值的作法，極其罕見於社會學調查資料分析。究其緣故，ML 或 EM 理論，對於許多社會學研究者來說，過於艱深難懂，因此，即使模型建構法具有諸多長處，反而，其社會學經驗研究的實用性不高（有關統計理論介紹，參見 Little and Rubin, 1987, 1989-90）。

近年來，Rubin（1987）提出一個「多重插補」（multiple imputation）的概念，主張運用各種方法插補和估計的數值，應該不限於一組，反之，研究者對於某一特定變項之遺漏值的處置，可以插補（或估計）一系列的數值。由於每一個遺漏值皆有對應的許多插補值或估計值，因此，研究者可以比較不同處置方法的差異，甚至估計插補的誤差，然後，進一步模擬估計值的分佈。可是，在實用的角度來看，由於多重插補法必須產生許多組群的插補值，然後重複模型分析，自然也就增加資料處理與分析的複雜性和成本。[13]

討論至此，我們已經介紹許多的遺漏值處置方法，那麼，這些眾多的處置方法之間的效用如何？過去以來，不少的研究者曾經試圖比較各種遺漏值處置方法之差異與效用（例如，翁彰佑、程爾觀，1991）。不過，正如同 Kromrey and Hines（1994）的看法，對於應用取向的研究者而言，不同處置方法之間的效用與差異，並不是關切的重心所在，而且，上列的研究顯示，對於各種遺漏值處置方法的效用，各家的評估結果甚爲分歧，如此益導致實用的研究者無所適從。

所以，我們贊同 Lessler and Kalsbeek（1992:229-233）的建議，第一，研究者必須謹記，「預防勝於治療」，所有事後的遺漏值處置策

13 請參見劉長萱、蔡政豐（1995）有關「多重插補」方法的討論和應用。

略，其效用永遠不會勝於「資料完整、毫無遺漏值」，因此，研究者最應集中心力於研究過程的設計以及資料蒐集。然而，社會學調查資料不可避免地可能遭遇遺漏值的難題，那麼，任何的事後處置措施，遠勝於鴕鳥式的視而不見作法——除非遺漏值的案例不多。當然，無論如何，實際的研究情境，才是最爲重要的考量。舉例而說，社會學調查資料中，甚少變項的屬性是連續性，而是以類別的屬性爲主，那麼，平均數替代或是迴歸方法就不可能派上用場。

參、一個遺漏值處置實例——熱卡插補

現在，我們就以實例說明，展示遺漏值處置的可能措施之一。在此，我們以「臺灣地區社會變遷基本調查（二期五次）」爲例，[14] 展示我們所採取的遺漏值處置策略。

首先，客觀而言，此一臺灣地區社會變遷基本調查（二期第五次）資料，可以說是相當地「乾淨」(clean)，也就是資料遺漏、流失的問題並不嚴重。不幸地，社會學研究者在許多實際的研究中，情境並不是如此「美好」——調查研究的人力、物力、經費、時間等，通常非常有限，因此，資料品質自然無法比擬中央研究院民族學研究所的社會變遷基本調查。

現在，研究主題或是研究所關切的重心是「上週工作時數」，研究者即將運用該調查中的文化價值問卷第四十九題資訊——「請問您上

14「臺灣地區社會變遷基本調查（二期五次）」係由瞿海源主持，中央研究院民族研究所執行完成之全臺灣地區社會調查計畫，詳見瞿海源主編《臺灣地區社會變遷基本調查計畫：第二期第五次調查計畫執行報告》(中央研究院民族學研究所，1994年)。

星期總共工作幾個小時？」。經過初步分析原始資料，我們得到表 1 的次數分配。

根據這個次數分配表，上週「有工作的受訪者」(1375 人) 當中，48 個個案屬於項目無反應 (不知道或拒答)。由於遺漏值的案例不多 (3.5%)，因此，研究者可以考慮只納入「完整的觀察體」進行直接分析。研究者不論採用成批抑或成對刪除方法刪除遺漏值個案，大體而言，分析結果差異不大。也就是說，以這個變項爲對象，研究者採用

表一 「上個星期總共工作時數」的次數分配

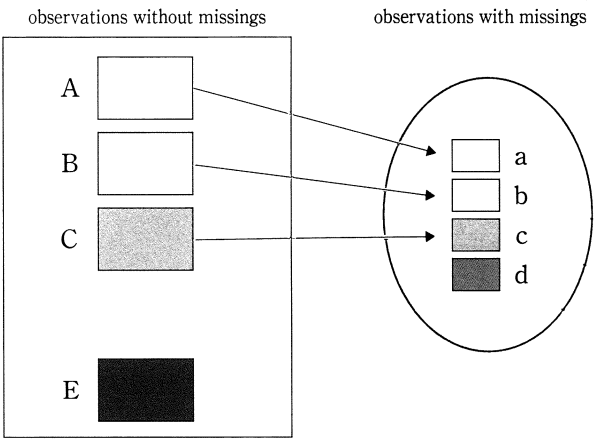
類別	Frequency	Percent
不適用	478	25.8
1- 25 小時	67	3.6
26- 30 小時	34	1.8
31- 40 小時	173	9.3
41- 50 小時	621	33.5
51- 60 小時	192	10.4
61- 70 小時	104	5.6
71- 80 小時	55	3.0
81- 90 小時	38	2.1
91-100 小時	26	1.4
101-110 小時	6	0.3
111-120 小時	7	0.4
121-130 小時	2	0.1
131-140 小時	1	0.1
141 小時以上	1	0.1
不知道	29	1.6
拒答	19	1.1
總計	1853	100.0

資料來源：「臺灣地區社會變遷基本調查 (二期五次)」文化價值組問卷第 49 題『您上個星期總共工作幾個小時？』。

成批或成對刪除而處置遺漏值，應該都是合理、可以接受的。不過，如果資料分析工作進一步進入多變項模型，那麼此一少量的遺漏值個案，仍有可能造成諸多問題。所以，我們考慮採取一些處置措施，試圖克服這個困難——以下採用熱卡插補法處置這些遺漏值。[15]

圖 1 就是我們採用的熱卡插補之基本架構。第一步工作，就是依照若干輔助變項的條件，將未出現遺漏值的觀察體劃分爲許多的「插補空格」。第二步工作，則是檢查遺漏值個案，從其所應對的插補空格中隨機找尋一個捐贈者，藉以替代插補遺漏值。過程中我們選擇若干與「上週工作時數」有關的輔助變項，用以形成插補空格。如果經過全部的找尋過程，仍有某些遺漏值個案無法取得捐贈者，那麼我們將減少輔助變項，重新形成插補空格，然後，重複熱卡插補找尋過程。

圖 1　熱卡插補的基本架構



15 在此，我們應用熱卡插補法的目的僅是「展示」的性質，也就是說，社會學調查資料中的遺漏值處置方法很多，並非限於熱卡插補，所以，研究者必須考量資料的性質、應用情境、以及研究者本身對於這些方法的偏好等因素。

現在，我們實際進行上述的熱卡插補工作。第一個工作，就是選擇輔助變項。根據社會學文獻，我們可以得知，一個人上週工作的時數，與其性別、年齡、教育程度、婚姻狀態、從業身份有關——譬如，男女的工作時數有別，已婚或未婚者所投入工作的時間也不同。我們以上週有工作的 1375 個觀察體爲對象，檢視其次數分配狀況，並且，比較不同性別、年齡、教育程度、婚姻狀態、從業身份的受訪者之工作時數，可以獲得結論支持選擇這些變項作爲輔助變項。圖 2 以盒型圖（box-and-whisker diagram）呈現這些次數分配，經過檢查評估，可以接受他們作爲輔助變項。

所以，我們透過性別、年齡、教育程度、婚姻狀況、以及從業身份等五個變項的「輔助」，將全部觀察體劃分爲一千個插補空格（2×5×5×5×4＝1000）。第二步工作的過程如下：

①第一個遺漏值個案，依照其輔助變項的條件，從一千個插補空格中找出對應的位置。

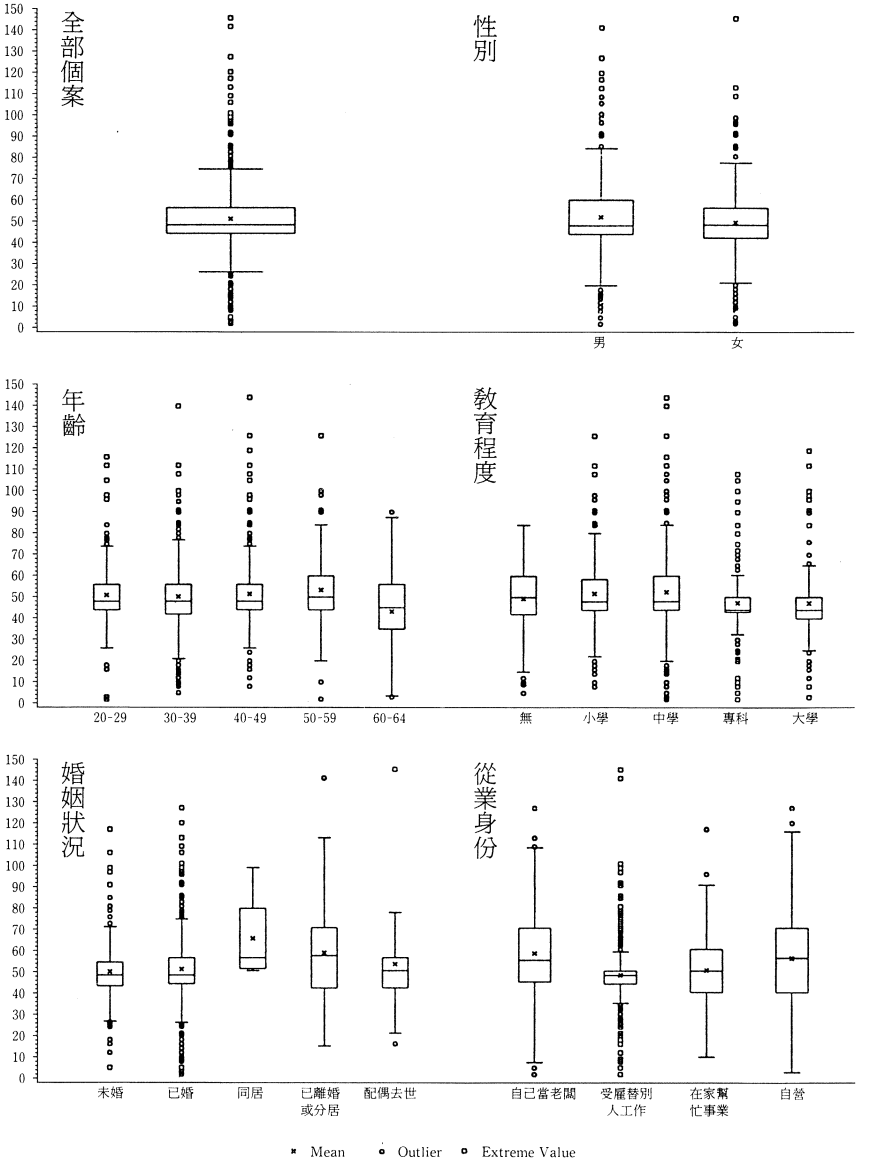
②在這個對應的插補空格中，隨機找出一個變項數值未流失的觀察體，將之作爲捐贈者，以其在該變項的數值插補替代遺漏值。

③處理第二個遺漏值個案，重複上述步驟。

④依順序處理全部 48 個觀察體。

上述步驟可以人工方式直接處理，不過，如果觀察體的規模很大，而且遺漏值案例不少，那麼人工處理就極不可能，勢必藉助於電腦協助。不幸地，目前的統計套裝軟體並未提供現成的程式以進行這些工作。本研究利用 SAS/IML 程式，撰寫一個 macro 執行上述隨機化熱卡插補的工作（此一 macro 列載於附錄一，可進行隨機化的熱卡插補替代工作。不過這個 macro 程式只能進行一輪的熱卡插補，所以，如果想要進行第二輪的插補工作，則依序重複執行這個程式）。

圖2 「上個星期總共工作時數」的分佈按若干特徵分



經過第一輪（cycle）的處置，48 個觀察體中，成功地插補替代 41 個遺漏值，不過，仍有 7 個觀察體的遺漏值未能插補。所以，針對這 7 個觀察體，我們進行第二輪的處置——只以年齡、教育程度、婚姻狀態、及從業身份等四個變項作爲輔助變項，形成插補空格，然後進行熱卡插補。結果，再插補 1 個遺漏值。因此，我們考慮第三輪處置，以教育程度、婚姻狀態、和從業身分等作爲輔助變項；最後，進入第四輪處置，僅以婚姻狀態和從業身份等兩變項輔助劃分插補空格，終於完全地將 48 個觀察體的遺漏值插補替代。表 2 列載四輪的插補處置

表二　遺漏值處置模擬結果：「上個星總共工作時數」的次數分配

類別	原始資料	第一輪處置	第二輪處置	第三輪處置	第四輪處置
不適用	478	478	478	478	478
1-25 小時	67	72	72	73	73
26-30 小時	34	35	35	37	37
31-40 小時	173	179	179	179	179
41-50 小時	621	637	638	639	639
51-60 小時	192	196	196	196	197
61-70 小時	104	109	109	110	110
71-80 小時	55	56	56	56	56
81-90 小時	38	39	39	39	39
91-100 小時	26	28	28	28	28
101-110 小時	6	6	6	6	6
111-120 小時	7	7	7	7	7
121-130 小時	2	2	2	2	2
131-140 小時	1	1	1	1	1
141 小時以上	1	1	1	1	1
總計	1805	1846	1847	1852	1853
遺漏值個案數	48	7	6	1	0

結果。

經過上述插補完成的「上週工作時數」變項，已經沒有遺漏值，可以進一步應用於單變項，甚或多變項分析。不過，研究者必須謹慎，畢竟這 48 個觀察體的變項數值乃是插補而來，所以，必要的時候，必須檢查其表現行爲是否顯著迥異於其他觀察體。

肆、討論

既然出現遺漏值的不完整資料乃是社會調查研究不可避免的難題，適當地有所作爲，仍是必要的。所以，社會學研究者，不應該忽略遺漏值處置的課題。近年來，社會學方法論在此正圖大力發展，至少兩個方向，已經日漸受到重視。

第一，遺漏值和樣本選擇偏誤的議題，不再完全被漠視，許多研究探討這些課題和實質理論（substance theory）之間的關係。例如，勞力市場社會學以及生涯流動的研究，對於觀察體耗損導致觀察值遺失的現象，逐漸重視它是否影響研究發現結果，甚至改變我們的理論認知。事實上，相對於傳統調查資料的遺漏值議題而言，近來，時貫性資料中由「censoring」、「truncation」所引發造成的不完整資料問題，似乎已經成爲社會學方法論的關注重心，尤其，事件史分析方法日益成爲社會學研究的重要工具之後，研究者對於不完整資料的議題更應賦予嚴正關切。16

16 有關 censoring 和 truncation 造成的不完整資料問題、以及相關的遺漏值處置之統計模型，請參見 Little and Rusibn （1987, Chap. 11）的深入說明。不過，大部份的社會學文獻則將之置於事件史分析方法的討論中加以處理，請參見陳信木、林佳瑩（1995）。

其次，針對個別的統計方法，已有許多深入的研究文獻，探討遺漏值處置問題，例如迴歸模型的遺漏值處置（Gourieroux and Monfort, 1981; Little, 1992b; Orme and Reis, 1991; Simon and Simonoff, 1986; Simonoff, 1988）、或是類別性資料分析的遺漏值處置（Ibrahim. 1990; Rindskopf, 1992; Siepman and Yang, 1994; van Buuren and Rijckevorsel, 1992; Winship and Mare, 1989）。此外，經過處置的遺漏值，如果研究想要探討其統計性質，諸如變異量估計或是漸近性（asymptotic），除了上述的「多重插補」途徑之外，最近，bootstrap sampling（參見 Stine, 1989-90）的抽樣模擬理論和技術，也可以有效運用於遺漏值的研究（Efron, 1994; Rubin, 1987）。事實上，這個途徑，對於社會學研究者而言，並不會太難於掌握，所以，應該有助於實際應用遺漏值處置技術。

參考文獻

伊慶春、蘇碩斌

1995 「無作答之分析：以公民容忍度爲例」，章英華、傅仰止、瞿海源主編，《社會調查與分析：社會科學研究方法檢討與前瞻之一》，頁 7-30。臺北：中央研究院民族學研究所。

李隆安

1995 「抽樣調查新方法的探討」，章英華、傅仰止、瞿海源主編，《社會調查與分析：社會科學研究方法檢討與前瞻之一》，頁 31-587-30。臺北：中央研究院民族學研究所。

洪永泰

1986 「抽樣調查中訪問失敗的問題」，〈思與言〉，23(6):65-71。

1989 「抽樣調查中訪問失敗問題的處理」，〈社會科學論叢〉，37:33-52。

1995 「抽樣調查中樣本代表性問題」，章英華、傅仰止、瞿海源主編，《社會調查

與分析：社會科學研究方法檢討與前瞻之一》，頁 7-30。臺北：中央研究院民族學研究所。

翁彰佑、程爾觀

1991 「隨機遺失資料插補法估計效用之比較」，〈中國統計學報〉，29(2):111-130。

陳信木、林佳瑩

1995 「勞工離職、轉業行爲之時間動態模型分析」，國科會專題研究計劃 NSC84-2412-H-004-002。

章英華、傅仰止、瞿海源（主編）

1995 《社會調查與分析：社會科學研究方法檢討與前瞻之一》。臺北：中央研究院民族學研究所。

楊文山、蔡瑤玲

1995 「實地調查中複查資料的結構模型分析：以臺灣地區社會意向調查爲例」，章英華、傅仰止、瞿海源主編，《社會調查與分析：社會科學研究方法檢討與前瞻之一》，頁 7-30。臺北：中央研究院民族學研究所。

劉長萱、蔡政豐

1996 「大型訪問調查的不完整取樣設計」，論文發表於〈第一屆「調查研究方法與應用」學術研討會〉，八十五年五月八日至十日，臺北，中央研究院調查研究工作室。

Afifi, Abdelmonem A. and Virginia Clark

1990 *Computer-Aided Multivariate Analysis*. Second Edition. New York: Van Nostrand Reinhold Co., Inc.

Afifi, Abdelmonem A. and R. M. Elashoff

1966 "Missing Observations in Multivariate Statistics—I. Review of the Literature." *Journal of the American Statistical Association* 61:595-604.

1967 "Missing Observations in Multivariate Statistics—II. Point Estimation in Simple Linear Regression." *Journal of the American Statistical Association* 62:10-29.

1969a "Missing Observations in Multivariate Statistics—III. Large Sample Analysis of Simple Linear Regression." *Journal of the American Statistical Association* 64:337-58.

1969b "Missing Observations in Multivariate Statistics—IV. A Note on Simple Linear Regression." *Journal of the American Statistical Association* 64:359-65.

Anderson, Andy B., Basilevsky, and Derek P. J. Hum

1983 "Missing Data: A Review of the Literature." Pp. 415-94 in *The Handbook of Survey Research*, edited by Peter H. Rossi, James D. Wright, and Andy B. Anderson. Orlando, Florida: Academic Press, Inc.

Arminger, Gerhard and Michael E. Sobel

1990 "Pseudo-Maximum Likelihood Estimation of Mean and Covariance Structures with Missing Data." *Journal of the American Statistical Association* 85(409):195-203.

Bailey, L., David W. Chapman, and Daniel Kasprzyk

1986 "Nonrespondense Adjustment Procedures at the U.S. Bureau of the Census." *Survey Methodology* 12:161-79.

Chapman, David W.

1976 "A Survey of Nonresponse Imputation Procedures." *American Statistical Association Proceedings of the Social Statistical Section* 1976(Part 1):245-51.

Chowdhury, Gopa

1991 "A Comparison of Covariance Estimators for Complete and Incomplete Panel Data Models." *Oxford Bulletin of Economics and Statistics* 53(1):83-93.

Cohen, Jacob and Patricia Cohen

1983 *Applied Multiple Regression/Correlation Analysis for the Behavior Sciences*. Second Edition. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

David, Martin, Roderick J. A. Little, Michael E. Samuhel, and Robert K. Triest

1986 "Alternative Methods for CPS Income Imputation." *Journal of the American Statistical Association* 81(393):29-41.

Dempster, A. P., N. M. Laird, and Donald B. Rubin

1977 "Maximum Likelihood from Incomplete Data Via the EM Algorithm." *Journal of the Royal Statistical Society, Series* B 39:1-38.

Efron, Bradley

1994 "Missing Data, Imputation, and the Bootstrap." *Journal of the American Statistical Association* 89(426):463-75.

Fay, Robert E.

1989 "Estimating Nonignorable Nonresponse in Longitudinal Surveys through Causal Modeling." Pp. 375-99 in *Panel Surveys*, edited by Daniel Kasprzyk, Greg Duncan, Graham Kalton, and M. P. Singh. New

York: John Wiley & Sons.

Ford, Barry L.

1983 "An Overview of Hot-Deck Procedures." Pp. 185-207 *in Incomplete Data in Sample Surveys*, vol. Volume 2, Theory and Bibliography, edited by William G. Madow, Harold Nisselson, Ingram Olkin, and Donald B. Rubin. New York: Academic Press.

Fuchs, Camil

1982 "Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data." *Journal of the American Statistical Association* 77(378):270-78.

Fuller, Carol H.

1974 "Weighting to Adjust for Survey Nonresponse." *Public Opinion Quarterly* 38(2):239-46.

Gelfand, Alan E. and Bradley P. Carlin

1993 "Maximum-Likelihood Estimation for Constrained- or Missing-Data Models." *The Canadian Journal of Statistics* 21(3):303-11.

Gilley, Otis W. and Robert P. Leone

1991 "A Two-Stage Imputation Procedure for Item Nonresponse in Surveys." *Journal of Business Research* 22(4):281-91.

Gourieroux, Christian and Alain Monfort

1981 "On the Problem of Missing Data in Linear Models." *Review of Economic Studies* 48(4):579-86.

Hartley, H. O. and R. R. Hocking

1971 "The Analysis of Incomplete Data." *Biometrics* 27:783-808.

Hayward, Mark D. and William R. Grady

1990 "Work and Retirement Among a Cohort of Older Men in the U.S., 1966-1983." *Demography* 27:337-56.

Ibrahim, Joseph G.

1990 "Incomplete Data in Generalized Linear Models." *Journal of the American Statistical Association* 85(411):765-69.

Jinn, J. H. and J. Sedransk

1989 "Effect on Secondary Data Analysis of Common Imputation Methods." *Sociological Methodology* 19:213-41.

Kalton, Graham and Daniel Kasprzyk

1982 "Imputing for Missing Survey Responses." *Proceedings of the Survey*

Research Methods Section, American Statistical Association.

1986 "The Treatment of Missing Survey Data." *Survey Methodology* 12:1-16.

Kalton, Graham and Leslie Kish

1984 "Some Efficient Random Imputation Methods." *Communications in Statistics, Theory and Methods* 13:1319-39.

Kim, Jae On and James Curry

1977 "The Treatment of Missing Data in Multivariate Analysis." *Sociological Methods and Research* 6(2):215-40.

Kromrey, Jeffrey D. and Constance V. Hines

1994 "Nonrandomly Missing Data in Multiple Regression: An Empirical Comparison of Common Missing-Data Treatments." *Educational and Psychological Measurement* 54(3):573-93.

Lee, Eun Sul, Ronald N. Forthofer, and Ronald J. Lorimor

1989 *Analyzing Complex Survey Data*. Newbury Park, California: Sage Publications, Inc.

Lepkowski, James M.

1989 "Treatment of Wave Nonresponse in *Panel Surveys*." Pp. 348-74 in Panel Surveys, edited by Daniel Kasprzyk, Greg Duncan, Graham Kalton, and M. P. Singh. New York: John Wiley & Sons.

Lessler, Judith T. and William D. Kalsbeek

1992 *Nonsampling Error in Surveys*. New York: John Wiley & Sons, Inc.

Little, Roderick J. A.

1988 "Missing-Data Adjustments in Large Surveys." *Journal of Business and Economic Statistics* 6(3):287-96.

1992a "Incomplete Data in Event History Analysis." Pp. 209-30 in *Demographic Applications of Event History Analysis*, edited by James Trussell, Richard Hankinson, and Judith Tilton. Oxford: Clearendon Press.

1992b "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87(420):1227-37.

Little, Roderick J. A. and Donald B. Rubin

1987 *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

1989-90 "The Analysis of Social Science Data with Missing Values." *Sociological Methods and Research* 18(2-3):292-326.

Little, Roderick J. A. and Hong-Lin Su

1989 "Item Nonresponse in *Panel Surveys*." Pp. 400-25 in Panel Surveys, edited by Daniel Kasprzyk, Greg Duncan, Graham Kalton, and M. P. Singh. New York: John Wiley & Sons.

Mandell, Lewis

1974 "When to Weight: Determining Nonresponse Bias in Survey Data." *Public Opinion Quarterly* 38(2):247-52.

Marini, Margaret Mooney, Anthony R. Olsen, and Donald B. Rubin

1980 "Maximum-Likelihood Estimation in Panel Studies with Missing Data." *Sociological Methodology* 11:314-57.

McArdle, John J.

1994 "Structural Factor Analysis Experiments with Incomplete Data." *Multivariate Behavioral Research* 29(4):409-54.

Mendoza, Jorge L.

1993 "Fisher Transformations for Correlations Corrected for Selection and Missing Data." *Psychometrika* 58(4):601-15.

Mislevy, Robert J.

1993 "Should "Multiple Imputations" Be Treated as "Multiple Indicators"?" *Psychometrika* 58(1):79-85.

Muthn, Bengt, David Kaplan, and Michael Hollis

1987 "On Structural Equation Modeling with Data That Are Not Missing Completely Random." *Psychometrika* 52(3):431-62.

Oh, H. Lock and Fredrick Scheuren, J.

1980 "Estimating the Variance Impact of Missing CPS Income Data." *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods.*

Oh, H. Lock, Fredrick Scheuren, J., and Harold Nisselson

1980 "Differential Bias Impacts of Alternate Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data." *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods.*

Orme, John G. and Janet Reis

1991 "Multiple Regression with Missing Data." *Journal of Social Service Research* 15(1-2):61-91.

Rindskopf, David

1992 "A General Approach to Categorical Data Analysis with Missing Data,

Using Generalized Linear Models with Composite Links." *Psychometrika* 57(1):29-42.

Rizvi, M. Haseeb

1983 "Hot-Deck Procedures: Introduction." Pp. 351-52 in *Incomplete Data in Sample Surveys*, vol. Volume 3, Proceedings of the Symposium, edited by William G. Madow, Harold Nisselson, Ingram Olkin, and Donald B. Rubin. New York: Academic Press.

Rovine, Michael J. and Mary Delaney

1990 "Missing Data Estimation in Developmental Research." Pp. 35-79 in *Statistical Methods in Longitudinal Research, Volume I: Principles and Structuring Change*, edited by Alexander von Eye. New York: Academic Press, Inc.

Rubin, Donald B.

1976 "Inference and Missing Data." *Biometrika* 70:41-55.

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

Sande, Innis G.

1979 "A Personal View of Hot Deck Imputation Procedures." *Survey Methodology* 5:238-58.

1982 "Imputation in Surveys: Coping with Reality." *The American Statistician* 36:145-52.

1983 "Hot-Deck Imputation Procedures." Pp. 339-49 in *Incomplete Data in Sample Surveys*, vol. Volume 3, Proceedings of the Symposium, edited by William G. Madow, Harold Nisselson, Ingram Olkin, and Donald B. Rubin. New York: Academic Press.

Siepman, Howard R. and Shie-Shien Yang

1994 "Generalized Least Squares Estimation of Multivariate Nonlinear Models with Missing Data." *Communications in Statistics. Theory and Methods* 23(6):1565-79.

Simon, Gary A. and Jeffrey S. Simonoff

1986 "Diagnostic Plots for Missing Data in Least Squares Regression." *Journal of the American Statistical Association* 81(394):501-09.

Simonoff, Jeffrey S.

1988 "Regression Diagnostics to Detect Nonrandom Missingness in Linear Regression." *Technometrics* 30:205-14.

Stine, Robert
1989-90 "An Introduction to Bootstrap Methods: Examples and Ideas." *Sociological Methods and Research* 18(2-3):243-91.
Van Buuren, Stef and Jan L. Van Rijckevorsel
1992 "Imputation of Missing Categorical Data by Maximizing Internal Consistency." *Psychometrika* 57(4):567-80.
Welniak, Edward J. and John F. Coder
1980 "A Measure of the Bias in the March CPS Earnings Imputation System." *American Statistical Association 1980 Proceedings of the Section on Survey Research Methods* 421-25.
Winship, Christopher and Robert D. Mare
1989 "Loglinear Models with Missing Data: A Latent Class Approach." *Sociological Methodology* 19:331-67.

附錄一 SAS Macro for Hot-Deck Imputation

```
/* ------------------------------------------------------------ *
 *  Randomized Hot-Deck Imputation of Missing Values            *
 *                                                by Hsinmu Chen *
 *    Four parameters need to be provided to use this sas macro.*
 *    SAS macro name: IMPUTE                                    *
 *    Parameters:      DATA, MISSVAR, ID, AUXVAR                *
 *                     DATA:    specify the sas dataset         *
 *                     MISSVAR: variable with missings to be imputed *
 *                     ID:      unique observation-ID           *
 *                     AUXVAR:  auxiliary variables             *
 *    Observations with missing-value will be imputed, in which the *
 *    donation-source could be identified by the FLAG variable. *
 *       Imputation FLAG variable denotes:                      *
 *         .           observation without missing              *
 *         0           observation with missing and not imputed *
 *         other value observation with missing but with imputed *
 *                     value donated from other observation     *
 *                     (ID value is specified)                  *
 * ------------------------------------------------------------ */;
OPTIONS CLEANUP NOMPRINT NOMLOGIC;

%MACRO IMPUTE (DATA =,      /* Specify the DATASET                    */
               MISSVAR =,   /* Variable with missing-values to be imputed */
               ID =,        /* Unique observation-ID                  */
               AUXVAR =)    /* Auxiliary-Variables                    */;
/*  Calculate # of Auxiliary-Variables  */
DATA _NULL_; ARRAY _XX_ &AUXVAR; DO OVER _XX_; NVAR + 1; END;
             CALL SYMPUT ('NVAR',TRIM (LEFT (PUT (NVAR,2.)))); RUN;
PROC  IML;
   /* ------------------------------ *
    * MISS:      matrix with missing-observations *
    * REF:       matrix without missing, for reference *
    * DONATED:   matrix to receive donation *
    * FLAG:      matrix with FLAG to denote imputation *
    * ------------------------------ */;
```

```
USE &DATA;
    READ  ALL VAR {&ID &AUXVAR}                    WHERE (&MISSVAR   = .)
          INTO MISS ( | COLNAME = ITEMNAME | );
    READ  ALL VAR {&ID &AUXVAR &MISSVAR}           WHERE (&MISSVAR ^= .)
          INTO REF;
CLOSE &DATA;
DONATED = REPEAT (.,NROW (MISS));
FLAG    = REPEAT (0,NROW (MISS));
  /* - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - *
   *  IMPUTATION MODULE                                              *
   *  This is the main part of this program to impute                *
   *  value of missing variable from the referent population.        *
   *  The module is consisting of three parts: 1) randomly           *
   *  sort the observations; 2) identify the corresponding           *
   *  cell based on &AUXVAR; 3) choose the first referent            *
   *  observation and donate the value, then repeat the next         *
   *  searching iteration.                                           *
   * - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - */;
  START IMPUTING;
    DO I = 1 TO NROW (MISS) ;
       REF = REF [RANK (RANUNI (REPEAT (0,NROW (REF)))),]
                  /* Randomly sort the referent population */;
       DO J = 1 TO NROW (REF) ;
       /*  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - *
        *  Iterates NROW (REF) times of searching to                  *
        *  search for donation from REFerent-matrix                   *
        *  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - */;
          DO K = 1 TO &NVAR;
            IF (MISS [I,K+1] ^=  REF [J,K+1] ) THEN GOTO NEXT;
          END;
          /* - - - - - - - - - - - - - - - - - - - - - - - - - - - - *
           *  Choose the corresponding cell based on                 *
           *  &AUXVAR.  Search the first observation                 *
           *  and donate the value as the imputed.                   *
           * - - - - - - - - - - - - - - - - - - - - - - - - - - - - */;
          DONATED [I]  = REF [J,&NVAR+2]   /* donation */;
          FLAG [I]     = REF [J,1]         /* flag ID  */;
          GOTO MATCHED;
       NEXT: END;
       NOMATCH: PRINT "No Matched Cell for :  &MISSVAR" ;
```

```
                    NOMATCH = MISS [I,] ;
                    VNAME    = {&ID &AUXVAR} ;
                    PRINT      NOMATCH [COLNAME = VNAME] ;
                    FREE       VNAME NOMATCH;
        MATCHED:  END  /* Next Missing-Observation */;
     FINISH;
     RUN IMPUTING;
   /*  - - - - - - - - - - - - - - - - - - - - - - *
    *  Write out the imputed matrix into dataset  *
    *  - - - - - - - - - - - - - - - - - - - - - - */;
    IMPUTE = MISS | | DONATED | | FLAG  /* vertically concatenate */;
    VARNAME = {&ID &AUXVAR &MISSVAR FLAG}   /* column label */;
    CREATE _IMPUTED FROM IMPUTE ( | COLNAME = VARNAME | );
    REPLACE;
    APPEND FROM IMPUTE;
    CLOSE _IMPUTED;
QUIT;
PROC SORT DATA = _IMPUTED; BY &ID;
PROC SORT DATA =&DATA   ; BY &ID;
DATA &DATA; UPDATE &DATA _IMPUTED; BY &ID; RUN;
%MEND IMPUTE;
```