國立臺灣大學生物資源暨農學院農藝學研究所

生物統計學組碩士論文

Division of Biostatistics, Graduate Institute of Agronomy

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

多重插補法在非完整資料統計分析上之應用

Application of Multiple Imputation on the Statistical Analysis of Incomplete Data

劉畢琳

Pi-Lin Liu

指導教授：劉清 博士

Advisor: Ching Liu, Ph.D.

中華民國 99 年 8 月

Aug，2010

國立臺灣大學碩士學位論文
口試委員會審定書

多重插補法在非完整資料統計分析上之應用

Application of Multiple Imputation on the Statistical Analysis of Incomplete Data

本論文係 劉畢琳 君（學號 R97621203）在國立臺灣大學農藝所生物統計學組完成之碩士學位論文，於民國 99 年 7 月 26 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

輔仁大學統計資訊學系暨應用統計所教授
商學研究所所長兼中華資料採礦協會理事長
謝邦昌 博士 謝邦昌

國立臺灣大學農藝學系(所)教授
謝英雄 博士 謝英雄

國立臺灣大學農藝學系(所)教授(指導教授)
劉　清 博士 劉清

誌　謝

光陰荏苒，匆匆的兩年就如同雨後的彩虹一般，一眨眼就過去了。在這兩年的生活中首先要感謝指導教授　劉清老師，老師的細心指導，諄諄教誨，以無比的耐心包容著我的不懂事以及許許多多的錯誤。也感謝　謝英雄老師和　謝邦昌老師於在百忙中擔任論文口試委員，提供許多寶貴的意見，讓論文得以順利完成。同時亦感謝生統組全體教師兩年來的教導和勉勵，讓我受益良多。

求學的兩年期間，感謝生統室諸位學長姐的照顧和指導，瑱芳、詩婷、雅純、西閔、蓓蓓、欣佩、博文、妙珊，還有一同擔任 TA 的欣怡學姐，更要感謝阿德和妍凌，讓我適應研究室的生活。還有一起學習的同學，亞靚、靖瑩和孟陽，在我疑惑時能幫助我了解，在我跟不上時能及時拉我一把，感謝有你們的陪伴，因為有了你們，使我的研究生活更加多采多姿。

還有在義光認識的曉婷、毓仁大哥和碧霜，在我緊張和不安時給予我支持與鼓勵，當然也不能忘了一直以陽光般燦爛笑容陪伴我的小老虎以及在背後默默支持我的托克比，感謝你們，伴著我度過每一個挫折，和我分享著每一個喜悅。

最後要感謝的是我的家人，在求學的路上始終給我百分之百的支持，在生活上給予我無微不至的照顧，讓我無後顧之憂的完成學業，讓我有所依靠，給我繼續下去的力量。而今，我將要告別求學生涯，懷著感恩的心，邁下人生的下一個旅程。

謹將這篇論文獻給伴我一路成長的老師、家人及同學們，因為有你們，才有現在的我，衷心的感謝你們，謝謝!

劉畢琳 謹誌
Pi-Lin, Liu
民國九十九年八月

摘　要

本文探討多重插補法在非完整資料統計分析上之應用。一般而言，在調查或收集資料時都會要求資料之完整，盡可能不要有缺失。但實際上，有許多情況下無法達到此要求，例如試驗單位的死亡，或者其它外在因素造成資料的闕無。當資料不完整時，會影響分析的效率，可能造成母體參數估計上的偏誤。所以需要推估缺失的資料點以形成完整之資料以利分析。

Rubin(1987)提出多重插補法，將每一個缺失值都代入 m＞1 個可能值，形成 m 套資料以供分析母體參數。本文以 SAS 9.1 User's Guild 中的例子作為完整資料，將資料隨機缺失 5%、10%、15%和 20%之後進行插補和分析，再與原始分析結果做比較，以了解多重插補法在使用上的成效。

模擬主要分為三部份：第一部份為可估計母體參數的統計分析方法，為迴歸分析和羅吉斯迴歸；第二部份為非估計母體參數的統計分析方法，為主成份分析、因素分析、鑑別分析、多變量分析和典型相關分析五種；第三部份是共變數結構的比較，為任意的共變數結構、混合對稱的共變數結構 、第一級自我迴歸的共變數結構和 Toeplitz 氏的共變數結構四種。

在第一部份得到的結果為在進行變數篩選時，迴歸分析會隨著缺失比例的增加而和完整資料所篩選出的結果差異漸增；而羅吉斯迴歸分析則無此情況，但此兩者的 P 值皆在缺失比例小時較能得到和完整資料相近的結果。在非估計母體參數的統計方法中，發現缺失插補後的分析結果最接近的為因素分析，其次為鑑別分析、多變量分析和典型相關三者，主成份分析和完整資料的差異最為明顯。共變數結構的模擬結果可發現結構若為任意的共變數結構、混合對稱的共變數結構和第一級自我迴歸的共變數結構三者並未改變，但是在 Toeplitz 氏的共變數結構中發現，若缺失比例較高時，可能會造成共變數結構的改變。

關鍵字：多重插補法、不完整資料、資料缺失、隨機缺失、馬可夫鏈蒙地卡羅

ABSTRACT

This paper investigates the application of multiple imputation on the statistical analysis of incomplete data. Many statistical analysis methods are designed and applicable only to complete data, and the incomplete data must be amended to meet the requirement.

Rubin (1987) proposed the method of multiple imputation by substituting m>1 possible values for each missing data. The resulting m sets of complete data are then subject to ordinary statistical analyses. The analysis results of these m sets of imputed completed data are combined together to provide for 5%, 10%, 15% and 20% missing proportions, and compared the analysis results with those of the original complete data.

Simulations in this paper were divided into 3 parts. The first is for the estimation of population parameters such as regression analysis and logistic regression. The second is for multivariate statistical analysis for multivariate normally distributed data. The third is about the covariance structures of multivariate data.

Results from the first part of simulation showed that the discrepancies of parameter estimates between complete data and incomplete data are proportional to missing proportion for regression analysis, but less obvious for logistic regression. Results from the second parts of simulations indicated that the factor analysis is most sensitive to missing proportion. Results from the third parts of simulations revealed that most of the variance structures studied in this paper are also robust to missing proportion.

Key words：Multiple Imputation, Incomplete Data, Missing Data, Missing at Random, Markov Chain Monte Carlo

目　　錄

表 目 錄

第一章 緒論

第一節 前言

雖然有些統計分析方法不需要完整資料即可進行分析，如變異數分析(analysis of variance)或是使用線性模式(linear model)進行統計分析，不但可以正確的進行統計分析，亦可得到良好的統計推論結果。然而現今的統計分析方法多為要求資料的完整才能夠進行分析，如迴歸分析或是因素分析等，若資料不完整(incomplete data)即無法進行統計分析。資料完整雖然是基本要求，卻是不容易達成的，造成資料不完整的原因有很多，最常見的是收集數據的過程中無法將資料完整的收集彙整(彭昭英，2009)。不論是在抽樣調查或是普查時，都有因為各種因素造成資料的缺失(missing data)。例如受試者填寫問卷時漏答(no response)等，都是造成資料不完整的原因。即便是在控制良好下的實驗也難免會產生缺失的情形，例如試驗單位的死亡，或者是田間試驗中突然的意外使資料缺無等(沈明來，2007)，也會造成資料的不完全。

由於這些資料點的缺失會直接影響到自由度的計算和統計檢定上的準確度，造成估計效率的降低，產生偏誤；對於推論統計值的解釋也造成一定的難度；更甚至使得某些統計分析方法不適用，影響層面相當廣大(彭昭英，2009)。目前針對這些缺失的資料點的處理方法共兩種，一為將整筆觀測值刪除，二是對這些缺失的資料點進行插補(imputation)。

一般而言不建議使用刪除法(deletion methods)，因為每一個資料點都帶有母體的資訊，在資料不完整的情況下，估計的效率已經降低，又再拋棄掉現有的資訊，將導致更大的偏誤；再者，有些研究的成本相當高，例如核磁共振等，無法重新實驗；或是樣本數相當稀少，拋棄掉任一筆資料都會使得分析結果有相當大的出入。

為了處理上述的問題，乃開始使用插補法；相對於刪去法，插補法將各遺失值帶入可能值以形成一組完整資料供統計分析之用，如此一來便可解決將帶有母體資訊的資料點刪除的缺憾。而插補法又分為單一插補法(single imputation)和多重插補法(multiple imputation)兩類。單一插補法顧名思義就是將各遺失值帶入一個可能值，形成一筆完整的資料，又可依其估算可能值的

方法再細分為均值取代法(mean substition)、簡易抽取法(simple hot-deck method)和迴歸估計法(regression-based method)等，其優點是簡單易操作，但有時估計不良時反而會造成更大的誤差；並且處理多變量的資料時也不易找到估算方法 (彭昭英，2009) 。而後 Rubins(1987)提出多重插補法，將每一筆缺失資料帶入 m>1 的估算值，形成 m 套資料以供分析，多重插補法的優點是能夠使不完整的資料重新形成一筆完整資料以供統計分析，並且能增加估計效率，使其統計推論的結果更加可信，因此也彌補單一插補法的不足之處。

本論文之目的是將多重插補法應用在各種不同的統計分析上，並比較不同統計分析方法下不同的缺失比例和完整資料分析結果之異同，將其分為三部份，一者為可估計母體參數的統計分析方法；第二為非估計母體參數的統計分析方法；最後是共變數結構的比較。

在下節中會先回顧前人的文章；而第二章針對多重插補法做詳細的介紹。在第三章中則是使用七種不同統計方法，將完整資料隨機遺失 5%、10%、15%和 20%，之後進行插補，將所得結果進行統計分析，以了解結果是否會和完整資料的分析結果產生差異；以及完整資料在隨機遺失插補後，是否會造成共變數矩陣的改變，最後一章則是結果與討論。

第二節　文獻回顧

多重插補法在 1987 年提出後，引起相當的討論，目前已廣泛應用於各領域中，在研究上，若遇到資料不完整的情況，有許多人選擇使用多重插補法做為解決資料不完整的方法，以形成完整資料進行統計分析。

雖然多重插補法並非適用於各種情況，在某些特定情況下，使用其它方法，如 EM 演算法，會更直接有效的得到完整資料(Schafer,1999)，但由於 EM 演算法無法求得概似估值變異數的不偏估計量（黃齡葦，2005），而無法進行假設檢定和統計推論，因此使用多重插補法仍不失為解決資料缺失的好方法之一。

李興南(2003)探討在多重插補法下的變異數估計和真實變異數的估計間的差異，證明在常態分佈下多重插補法參數估計量的變異數估計量是不偏的估計量。

楊棋全(2004)則針對指數分布和韋伯分布下遺失值的處理方法中進行比較，考慮四種不同遺失比例 20%、50%和 80%，樣本數 10、20 、40、60、80 和 100，搭配不同的插補次數，來比較不同插補法的優缺點。研究結果發現，多重插補法並非適用於指數分布和韋伯分布，並且插補次數應該略大於 Rubin 所建議的 5-12 次，增加至 8-20 次才能獲得較穩定的 MSE 值。而楊棋全發現在指數分布下，該估計量並非不偏，故多重插補法中的變異數估計量並非對任何分布都適用 (楊棋全，2004) 。

黃齡葦(2005)比較了多重插補法和拔靴法(bootstrap method)及資料擴增法(data augmentation)三者在區間估計上的成效，目的為證明多重插補法的參數估計能力可以達到和拔靴法一樣的效果。

黃齡葦(2005)比較了三種不同的資料型態，第一組資料結為小樣本資料，共有 20 個樣本，包含兩變數，且變數皆成常態分佈；第二組為中樣本，樣本數為 150，有三個變數；最後一組為大樣本資料，500 個樣本，五個變數。然後隨機遺失掉 10%、30%和 50%。再用這些不完整的資料進行三種方法的信賴區間估算，以比較其成效。

比較的結果為多重插補法的估計能力雖然會受到缺失比例高低和樣本大小的影響，但相對於其他兩種方法，多重插補法估計區間的估計是較優良的，沒有出現明顯的偏誤（黃齡葦，2005）。

本文採用類似的方法，以 SAS 9.1 User’s Guild 中的例子為完整資料，然後隨機遺失特定比例，再比較和完整資料的差別。

第二章 統計方法

調查或收集而來的資料若有缺失，在缺失比例小於 5%的情況下，可將不完整的觀測值刪除進行統計分析，仍然可得到良好的分析結果(Schafer，1997)；但是當缺失比例很大時，刪去法會拋棄過多的資訊，產生估計上的偏誤。此時，插補法便提供另一種解決方式，將每一個缺失位置都代入可能值，形成完整資料已進行分析。而本章所介紹的多重插補法，由 Rubin(1987)提出，其基本概念是將每一個缺失值都代入 m＞1 個可能值，形成 m 套資料以供分析母體參數，而 m 通常介於 3 到 10 之間，並不會使用太大的插補套數(Schafer，1999)。

多重插補法的在使用上的優點為可將有缺失的資料形成完整資料進行分析，增加估計效率，並且使其統計推論的結果更加可信。缺點則是產生多重插補的資料相較於其他方法複雜，並且由於產生多套資料，需要較多的空間來儲存資料，多重插補後的資料分析較複雜，簡而言之，多重插補法的缺點為在產生、儲存和分析上都較單一插補法繁複(Rubin，2004)。

一般使用多重插補法有三步驟：第一步將各缺失值代入 m＞1 個可能值，形成 m 套完整資料，一般情況下，採用 m=5 套資料已足夠，但當資料缺失的比例越高，所需形成的套數也越多；第二步是使用這 m 組資料進行一般對完整資料的統計分析；最後將所得估計值加以結合，以獲得插補後的參數估計值 (黃齡葦，2005) 。

第一節　多重插補法之前提假設

多重插補法的前提假設有二：一為資料中的缺失必須為隨機遺失(missing at random; MAR) ；二為資料的型態為多元常態分布(SAS，2002)。

假設 Y 為有 p 個變數 n 個觀測值的完整資料，形成一個 n×p 矩陣，如 $\begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{pmatrix}$，將其中每一個位置稱為資料點，故 n×p 矩陣中共有 n×p 個資料點，而其中的元素又可分為 Y_{obs} 和 Y_{mis} 兩類，Y_{obs} 為已觀測到的資料點；Y_{mis} 為缺失的資料點，隨機遺失意指每個缺失值都只和 Y_{obs} 有關，並不會和 Y_{mis} 產生關係。

進一步，又令 R 亦為一個 n×p 矩陣，矩陣中各位置對應到 Y，以 1 表示

Y_{obs}，以 0 表示 Y_{mis}，在隨機遺失的情形下，可表示為

$$Pr(R|Y_{obs}, Y_{mis})=Pr(R|Y_{obs}) \quad (1)$$

亦即資料的缺失只和已觀察到的值相關，各遺失值間是獨立的，不會相互影響。

由於在 SAS 中，多重插補法是使用馬可夫鏈蒙地卡羅法(Markov chain Monte Carlo；MCMC)而來，因此會要求缺失的資料型態為多元常態分布(SAS，2002) 。馬可夫鏈（Markov chain）是一系列隨機變數的集合，其分布中第 n 個元素會與第 n-1 個元素相關，即每個元素會受前一個元素影響，而且僅和前一個元素有關；而馬可夫鏈蒙地卡羅法即是利用馬可夫鏈為基礎進行模擬，為一種數值研究方法，利用隨機取樣的方式來模擬隨機變數的機率分配，進而取得一些重要參數。

利用貝氏定理 $p(\theta|y)=\frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$ 可求得未知的參數，在此處馬可夫鏈蒙地卡羅法及利用貝氏定理的事後機率概念，進行模擬以求得參數(SAS，2002)。方法為先對欲估計的母體參數 θ 給定先驗分配(prior distribution)，接著再算出 θ 的事後分配

$$p(\theta|Y_{obs,}Y_{mis}) \equiv p(\theta|Y_{obs})=c\times p(\theta)\times f(Y_{obs}|\theta) \quad (2)$$

其中，c 為常數，$p(\theta)$為參數 θ 的先驗分配，$f(Y_{obs}|\theta)$為可觀測資料的機率密度函數(黃齡葦，2005) 。

之後分為兩步驟遞迴而至得到一穩定不變之數為止，第一個步驟為插補(imputation step；I-step)，在給定的平均向量(μ)和變異數矩陣(Σ)中，I-step 要將每一個觀測值中的缺失值分別模擬，也就是說，從 $p(Y_{mis}|Y_{obs,}\theta^{(t)})$抽取出 $Y_{mis}^{(t+1)}$，其中，θ 為欲估計的母體參數值，$\theta^{(t)}$則是 θ 的估計值；在 SAS 中會利用 EM 演算法求得(μ，Σ)的最大概似估計量(MLE)做為起始值，也就是第一次進行 I-step 時所需要的平均向量和變異數矩陣(SAS，2002) 。

第二步為事後機率步驟(posterior step；P-step)則是在 I-step 後會形成一組完整資料，在完整資料下，由 $p(\theta|Y_{obs,}Y_{mis}^{(t+1)})$抽出 $\theta^{(t+1)}$，所得到的參數估值 $\theta^{(t+1)}$

為進行下一次 I-step 所需之平均向量(μ)和變異數矩陣(Σ)。如此重複多次即得馬可夫鏈

$(Y_{mis}^{(1)}, \theta^{(1)})$，$(Y_{mis}^{(2)}, \theta^{(2)})$，…，

會收斂至 $p(Y_{mis}, \theta | Y_{obs})$，所得之值即可將其插補至資料中 (SAS，2002) 。

第二節　資料整合分析的方法

假設 Q 為欲估計的參數，而令 $\hat{Q} = Q(Y_{mis}, Y_{obs})$ 為完整資料下得到的統計量，令 $U = U(Y_{mis}, Y_{obs})$ 為完整資料下得到的樣本變異數，在常態分布的假設下可得

$$\frac{(\hat{Q} - Q)}{\sqrt{U}} \sim N(0,1) \tag{3}$$

而在不同情況下，可進行不同的轉換，如當欲估計的參數為勝算比(Odd's ratio)時，可進行對數轉換；當樣本數小時，則為學生氏 t 分布 (Schafer，1999) 。

定義 $Y_{mis}^{(1)}$，$Y_{mis}^{(2)}$，$Y_{mis}^{(3)}$，…，$Y_{mis}^{(m)}$ 為 m>1 套相互獨立的插補值，從而計算出估計值和變異數分別為

$$\hat{Q}^{(l)} = Q(Y_{mis}^{(l)}, Y_{obs}) \tag{4}$$

和

$$U^{(l)} = U(Y_{mis}^{(l)}, Y_{obs}) \tag{5}$$

其中 l=1,2,…,m。再據此算出整體平均

$$\bar{Q} = \frac{\sum \hat{Q}^{(l)}}{m} \tag{6}$$

和組間變異 (between-imputation variance) $B = \frac{\sum (\hat{Q}^{(l)} - \bar{Q})^2}{(m-1)}$，組內變異 (within-imputation variance) $\bar{U} = \frac{\sum U^{(l)}}{m}$，總變異(total variance)和

$$T = (1 + \frac{1}{m})B + \bar{U} \tag{7}$$

因此可得

$$\frac{(\hat{Q}-Q)}{\sqrt{T}} \sim t_v \tag{8}$$

其自由度

$$v = (m-1)\left[1+\frac{\bar{U}}{(1+m^{-1})B}\right]^2 \tag{9}$$

(Schafer,1999)。計算 $r = \frac{(1+m^{-1})B}{\bar{U}}$，代表缺失造成相對增加的變異（the relative increase in variance due to missing data）。據此又可得統計量

$$\hat{\lambda} = \frac{r+2/(v+3)}{r+1} \tag{10}$$

代表缺失訊息的比例(the fraction of missing information)。r 和 λ 代表著欲估計的參數 Q 受到缺失的影響大小，假設一筆資料的缺失比例為 10%，表示資料共遺失 10%個資料點，但對每一個欲衡量的變數而言，受到遺失的影響可能不只有 10%，可能超過 10%；因此 r 和 λ 這兩者指標是越小越好。(SAS，2002)。

分析時為了決定插補次數，可由相對效率(relative efficiency)計算，

$$RE = (1+\frac{\lambda}{m})^{-1} \tag{11}$$

表 1 列出不同缺失訊息的比例（λ）和插補次數（m）下所計算出的相對效率，可看出當缺失訊息較小時，不需要太大的插補次數即可得到較大的效率。(SAS，2002)，若缺失訊息的比例不太大時，一般使用 5-10 次即可(Schafer，1999)。

表 1. 相對效率

Table 1. Relative efficiency (SAS，2002)

	λ				
m	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

第三節 比較資料結構的方法

在非估計母體參數的統計方法中，皆採用多變量分析方法，根據不同的統計分析方法，採用不同的方法衡量缺失和完整資料的差距，在主成份分析中，將對各新變數有顯著貢獻的因子篩選出來，在本文中，若係數大於 0.4 或小於-0.4 即稱為有顯著貢獻，其後比較完整資料篩選出的因子和缺失後插補的因子是否相同，相同的個數越多，代表和完整資料越相近，缺失後插補的成效亦越好。而在因素分析中，將資料經過最大變異法(varimax)的轉換，使得因素荷重向 0 或 1 靠近，當因素荷重大於 0.7 時，代表該因子對因素具有解釋能力，比較完整資料中，對各因素有解釋能力的因子，選出的因子相同數越多，代表和完整資料越相近，缺失後插補的成效越好。

在鑑別分析中，則是由錯分的個數和錯分率(misclassification rate)來判斷，將完整資料下的錯分個數及錯分率和缺失後的錯分個數及錯分率相比較，數字越接近代表和完整資料越相近。而多變量變異數分析的比較方式為以 P 值和 F 值作判斷，和完整資料越接近者代表和完整資料越相近，缺失後插補的成效越好，若有達到顯著，則繼續進行多重比較(multiple comparison)。典型相關分析先判斷典型相關係數和 P 值的大小，是否會接受或拒絕虛無假設，即兩典型相關變數間沒有相關性存在；之後再判斷各典型變數的組成，是否和完整資料相同，若典型相關係數、P 值和各典型變數的組成和完整資料相似，則可判斷其和完整資料相似；若僅有典型相關係數和 P 值和完整資料相似或是僅各典型變數的組成和完整資料相似，則判斷其和完整資料的相似度中等；若皆不相似，則判斷其與完整資料差異很大，完整不相似。

第四節 比較共變數結構異同的方法

在共變數結構中，先假設一共變數結構(Σ)，並產生一筆有 5 個變數，觀測個數(N)為 100 的資料，隨機缺失特定比例且進行多重插補後，判斷差補補後資料的共變數結構是否有所改變。

判斷方法為定義虛無假設 H_0：$\Sigma = \Sigma_0$ 後，計算統計量

$$L = v(\ln|\Sigma_0| - \ln|S| + trS\Sigma_0^{-1} - p) \tag{12}$$

和

$$L1 = \left\{1 - \frac{1}{6(N-1)}\left[2p + 1 - \frac{2}{p+1}\right]\right\} \tag{13}$$

其中 S 為要計算所得的矩陣， v 為 S 的自由度， p 則為變數數目，N 為觀測個數。$L1$ 會服從卡分自由度為$\frac{1}{2}$p(p+1)的分布，若 $L1$ 大於 $\chi^2_{\alpha,\frac{1}{2}p(p+1)}$，則拒絕虛無假設，代表兩共變數矩陣不相同(Morrison，1990)。

第三章　資料模擬和分析

本章資料模擬所採用例子由 SAS 9.1 User's Guild 中的例子為完整資料，其後再隨機缺失 5%、10%、15%和 20%的資料點，將此四筆有缺失的樣本分別以多重插補法進行插補，各形成五套完整資料進行統計分析，之後再與完整資料比較分析結果，判斷其差異。

可分為三個部份，第一部份是可估算母體參數的分析方法，為迴歸分析和羅吉斯迴歸，其中又再採取兩種方法進行母體參數的估計，第一種為將多重插補法形成的五套資料以第二章第二節所提之方法進行整合分析，稱為方法 A，第二種為插補後，將所形成五套資料中各觀測值平均而形成一套完整資料後再進行分析，稱為方法 B。

第二部份為非估算母體參數的分析方法，為主成份分析、因素分析、鑑別分析、多變量分析和典型相關分析五種，皆為多重插補法後平均五套資料形成一套完整資料後進行分析，再與完整資料進行比較。

第三部份則是共變數結構的比較，先設定一共變數結構，之後隨機產生一筆個數為 100 的多元常態分布資料，並且使這筆資料的共變數結構與設定之共變數結構沒有差異，令此資料為完整資料，之後再由完整資料隨機缺失 5%、10%、15%和 20%，插補之後平均五套資料形成一筆完整資料，再求得缺失後插補之共變數結構和原設定的共變數結構之異同。

第一節　迴歸分析(regression analysis)

迴歸分析是用來分析兩個或兩個以上變數間的關係，在實務上非常重要的分析工具，其方法為根據已知且可控制的有關變數的變數值去估計或預測另一變數的變數值。迴歸分析的主要目的是在建立變數與變數間的統計關係，利用此統計模型去做預測 (Moore & McCabe，2003) 。

本文採用 SAS 9.1 User's Guild 中，Example 61.1：aerobic fitness prediction 例子為完整資料，探討耗氧量和其它變數間的關係。變數為年齡(age) 、體重(weight) 、耗氧量(oxygen) 、跑步時間(run time) 、 靜止時脈搏數(rest

pulse) 、跑步時脈搏數(run pulse) 、最大脈搏數(max pulse) 七個變數，31 筆觀測值，一共 217 個資料點，分別隨機缺失 10 個、21 個 、31 個和 42 個資料點後分別以第三章所提之方法 A 及 B 分別進行迴歸分析篩選出顯著的變數，所得之結果列於表 2。

表 2. 迴歸分析之變數篩選結果(P 值)

Table 2.Variable selection result (P-value) from regression analysis

	complete	5%		10%		15%		20%	
		A	B	A	B	A	B	A	B
intercept	<.0001**	<.0001**	<.0001**	0.0007**	<.0001**	0.0727	<.0001**	<.0001**	<.0001**
age	0.0322*	0.0286*	0.0292*	0.4784	0.3278	0.9466	0.5905	0.0120*	<.0001**
weight	0.1869	0.1848	0.1636	0.5625	0.3845	0.8210	0.4380	0.1646	0.0526
run time	<.0001**	<.0001**	<.0001**	<.0001**	<.0001**	<.0001**	<.0001**	<.0001**	<.0001**
run pulse	0.0051**	0.0066**	0.0080**	0.2759	0.1027	0.4740	0.4287	0.9328	0.5889
rest pulse	0.7473	0.9724	0.9543	0.6682	0.3395	0.4872	0.2974	0.4102	0.0195*
max pulse	0.0360*	0.0484*	0.0423*	0.5317	0.2887	0.6367	0.8407	0.4433	0.0170*

由表 2 可以看出，在完整資料下進行迴歸分析可以篩選出截距項和四個變數，分別為年齡、跑步時間、跑步時脈搏數和最大脈搏數。在缺失值比例為 5%時，方法 A、B 和完整資料三者所篩選出的變數完全相同，而 P 值的差距甚微；而缺失值比例為 10%時，年齡、最大脈搏數兩個變數並沒有被篩選出來，而雖然方法 A 和 B 篩選出的變數相同，但兩者的 P 值差距稍有增加。缺失值比例到達 15%時，方法 A 僅篩選出跑步時間此一變數，方法 B 則篩選出截距項 、跑步時間兩者，可看出三者之間的差距會因為缺失值比例的增加而使得差距隨之增加。最後在缺失值比例為 20%的情況下，方法 B 較方法 A 多篩選出了靜止時脈搏數和最大脈搏數，但靜止時脈搏數並不是完整資料所篩選出來的結果；而兩種方法都沒有篩選出跑步時脈搏數。

由表 3 可以看出，在缺失 5%、10%、15%和 20%時，缺失值對各因子的影響。在缺失 5%時，對截距項、年齡、跑步時脈搏數和最大脈搏數影響較大，

對於這些變數而言，缺失的比例已經超過 30%，但從相對效率來看，各變數皆達到 90%，表示相對於沒有缺失時，缺失 5%的資料估計效率仍可達到 90%；在缺失 10%時，受到影響的變數略有不同，截距項、體重、跑步時脈搏數和最大脈搏數，但相對效率仍有 90%以上。

然而，在資料缺失比例達到 15%和 20%時，對各因子的影響都很巨大，皆達到 30%以上，並且相對效率亦降低，在缺失 15%時截距項的相對效率已降低至 87.3%；而在缺失 20%時，僅體重的相對效率由達到 90%以上，但其餘仍可維持 85%以上，若需要更高的相對效率，可參考表 1，增加多重插補的套數，以求得更高的估計效率。

表 3. 迴歸分析中缺失訊息的比例和相對效率

Table 3. The fraction of missing information and relative efficiency from regression analysis

	missing 5%		missing 10%		missing 15%		missing 20%	
parameter	λ	RE	λ	RE	λ	RE	λ	RE
intercept	0.357	0.933	0.312	0.941	0.727	0.873	0.610	0.891
age	0.362	0.932	0.171	0.967	0.491	0.911	0.608	0.892
weight	0.191	0.963	0.448	0.918	0.416	0.923	0.514	0.907
run time	0.171	0.967	0.122	0.976	0.420	0.922	0.638	0.887
run pulse	0.323	0.939	0.421	0.922	0.451	0.917	0.769	0.867
restpulse	0.206	0.961	0.218	0.958	0.343	0.936	0.720	0.874
max pulse	0.341	0.936	0.332	0.938	0.393	0.927	0.788	0.864

第二節　羅吉斯迴歸(logistic regression analysis)

迴歸分析是分析一個應變數與一個或多個解釋變數之間的關係，其中，欲探討的應變數為離散型(discrete)變數時，便使用羅吉斯迴歸作為分析的工具。不同於簡單迴歸，羅吉斯迴歸是使用在反應變數 Y 為二元變數(0 或 1)而解釋變數 X 仍為連續變數的情況。其模型可寫成

$$\text{logit}(p)=\log\left(\frac{p}{1-p}\right)=\alpha+\beta X \qquad (14)$$

其中 p=Pr(Y=1|X=x)，α 為截距項，β 為斜率，同時也是要估計的母體參數(Agresti，1990)。在進行假設定和統計推論時，會設立虛無假設為 $\beta_i=0$，i=1,2….n，表示欲檢定解釋變數 X 對於反應變數 Y 是否有影響力，因此，當檢定結果為拒絕虛無假設時(P 值小於 0.05)，便可宣稱解釋變數 X 對反應變數 Y 有顯著的影響；反之，則表示無影響。

本文使用 SAS 9.1 User's Guild 中 Example 42.1：stepwise logistic regression and predicted values 的資料為完整資料。實驗共有 27 位受試者，目的為欲了解病人特性對癌症治療是否有緩解效果，有緩解為 1，沒有緩解為 0，解釋變數為 cell、smear、infil、li、blast 和 temp 六者，此六者為多元常態分布，故共有 162 個資料點。隨機缺失 8 個、16 個、24 個和 32 個後，如同第一節使用方法 A 和方法 B 進行比較，是否能篩選出相同的變數，所得之結果列於表 4。

由表 4 可以看出，完整資料時並沒有篩選出任何變數，在進行多重插補後，不論是使用方法 A 或方法 B 亦沒有篩選出任何變數。推測可能的原因是因為在完整資料所得到的 P 值都偏大，並沒有靠近臨界值，因此不顯著的變數仍然維持不顯著。對照表 2，當 P 值相當顯著時(＜0.0001)， 插補所得之 P 值亦相當顯著(截距項和跑步時間)，而像年齡和體重兩變數的 P 值較靠近臨界值(分別為 0.0322、0.0360)，在插補過後的變動就較為明顯。

從表 5 來看，缺失比例對各變數的影響都較表面上來的大，在缺失比例為 5%和 10%時，相對效率仍可達到 90%以上，而資料隨著缺失比例的增加，所得到的相對效率會漸漸降低，在資料缺失比例達到 20%時，其估計效率僅 li 和 blast 兩者有達到 90%，其它變數則不到 90%，但仍可維持 85%以上，表示和沒有缺失的資料相比，估計效率仍有 85%以上。而在缺失對各變數的影響上，除了在缺失資料為 10%時，其他三筆資料所呈獻的影響都偏大，特別是在缺失比例為 5%時，對大部份變數的影響即達到 40%以上。和迴歸分析相比較，羅吉斯迴歸似乎更容易受到缺失值的影響。

表 4. 羅吉斯迴歸之變數篩選(P 值)

Table 4. Variable selection result (P-value) from logistic regression analysis

parameter	complete	5%		10%		15%		20%	
		A	B	A	B	A	B	A	B
intercept	0.4152	0.3593	0.2323	0.2802	0.2371	0.1347	0.0690	0.7003	0.0900
cell	0.6062	0.4669	0.3862	0.7648	0.6261	0.9030	0.6354	0.6812	0.8690
smear	0.7392	0.5122	0.4352	0.9676	0.7979	0.9790	0.7028	0.6341	0.8142
infil	0.7507	0.5035	0.4126	0.9598	0.7817	0.9679	0.6773	0.6458	0.8841
li	0.0955	0.2577	0.1270	0.2503	0.1690	0.5281	0.7639	0.5405	0.2015
blast	0.9471	0.4879	0.3396	0.7481	0.5798	0.4011	0.1492	0.8845	0.2631
temp	0.1957	0.3073	0.1491	0.2265	0.1826	0.1142	0.0445	0.7806	0.1257

表 5. 羅吉斯迴歸下缺失訊息的比例和相對效率

Table 5. The fraction of missing information and relative efficiency from logistic regression analysis

parameter	missing 5%		missing 10%		missing 15%		missing 20%	
	λ	RE	λ	RE	λ	RE	λ	RE
intercept	0.259	0.951	0.116	0.977	0.532	0.904	0.712	0.875
cell	0.515	0.907	0.199	0.962	0.773	0.866	0.779	0.865
smear	0.509	0.908	0.259	0.951	0.778	0.865	0.685	0.879
infil	0.501	0.909	0.264	0.950	0.771	0.866	0.661	0.883
li	0.466	0.915	0.286	0.946	0.556	0.900	0.546	0.902
blast	0.184	0.964	0.096	0.981	0.283	0.946	0.427	0.921
temp	0.407	0.925	0.106	0.979	0.332	0.938	0.598	0.893

第三節　主成份分析(principal component analysis)

主成份分析為多變量分析的方法之一，可作為其它分析的過渡性媒介，如因素分析、集群分析，多元迴歸分析等等。將原來多個變數經過線性組合後形成新的變數，即假設原來的變數為 $X_1,X_2,\ldots,X_p$，要將這 p 個變數經過線性組合後變成 $P_1,P_2,\ldots P_P$，P 個新變數，稱為主成份 1，主成份 2，...主成份 P。目的是要將龐大的資料縮減，找出之間的關連 (Johnson，2007) ，所以，一般線性組合後的新變數不會取 P 個，而是由其累積解釋變異百分比來決

定，方法為先計算線性組合後新變數的變異數，為$Var(Y_i) = e_i'\Sigma e_i = \lambda_i$，其中Σ代表共變數矩陣，$e_i$代表第 i 個變數的特徵向量（eigenvector），λ_i代表第 i 個變數的特徵值(eigenvalue)，若算出的變異越大代表能解釋資料中的變異越多，之後將變數的解釋變異百分比由大到小排列並相加以求得累積變異解釋比例，若達到 80%以上，就取其個數為新的變數個數(Johnson，2007)。

本文所使用 SAS 9.1 User's Guild 中 Example 58.2：crime rates 的資料為完整資料，內容為 1977 年美國五十州的犯罪率，以七個指標呈現，分別為謀殺(murder)、強姦(rape)、搶劫(robbery)、 暴力攻擊(assault)、入室竊盜(burglary)、偷竊(larceny)以及偷車(auto theft)，共 350 個資料點。之後再隨機缺失 17 個、35 個、52 個和 70 個資料點，進行多重插補後，將五套資料平均形成完整資料後進行主成分分析，和完整資料所得結果做比較。所得資料整理於表 6。

由表 6 可以看出，不論缺失比例為多少，平均數和完整資料所計算出來的數值並沒有太大差異。值得注意的是直覺上當缺失比例越高，會和完整資料相差越遠，但由此表中可以看出，結果並非如此。以謀殺這個指標來看，完整資料的平均為 7.44(單位：百萬分之一人)，隨著缺失比例的增加，其平均和完整資料平均的差距分別為 0.2、0.1、0.22 和 0.13，由此可知，插補後的分析結果和完整資料的分析結果，兩者的差距不一定是和缺失比例的高低呈現正向關係。

主成份分析時，取線性組合變數個數為其累積解釋百分比達到八成以上即可，由表 5、表 6、表 7 可以看出完整資料和缺失 5%、10%時，需取三條線性組合為新變數，而 15%、20%時則在第二條線性組合即達到要求(皆為 0.820)。

表 7 列出各資料分析所得之第一條線性組合(主成份 1) 而成的新變數，標記為 a 代表該因子達到對新變數的貢獻是正向顯著的(大於 0.4)，而標記為 b 的數字代表該因子對新變數的貢獻是負向顯著(小於-0.4)。在完整資料下各因子對此新變數的貢獻差距不大，雖然只有強姦和入室竊盜兩者達到正向顯著，但強劫和暴力攻擊也十分接近顯著(皆為 0.397)，造成缺失插補後可能造

成達到正向顯著，可看出資料缺失比例越高，各因子對新變數的貢獻差異越大。

表 8 列出各資料分析所得第二條線性組合(主成份 2) 而成的新變數，同樣的標記為 a 是正向顯著貢獻，標記為 b 是反向顯著貢獻。由表五可以看出，在第二條線性組合中，在不同比例的缺失值下，有顯著貢獻的為謀殺、竊盜和偷車三因子，但是貢獻的方向卻會改變，並非一致。

而表 9 是第三條線性組合(主成份 3)而成的新變數，可看出在此新變數下，不論是完整資料或各缺失比例下有貢獻的因子皆相同，且方向也一致。

表 6. 完整資料與各缺失比例所得之平均和標準差

Table 6. Mean and standard deviation of different missing data sets

variable	complete	5%	10%	15%	20%
murder	(7.44,3.87)	(7.46,3.85)	(7.45,4.09)	(7.22,4.01)	(7.31,4.04)
rape	(25.73,10.75)	(26.31,10.57)	(26.14,10.51)	(25.81,10.18)	(25.19,9.98)
robbery	(124.09,88.35)	(127.75,91.58)	(127.78,92.21)	(124.58,86.89)	(124.26,87.32)
assault	(211.30,100.25)	(209.40,103.15)	(211.52,101.77)	(212.60,99.22)	(209.64,103.49)
burglary	(1291.90,432.46)	(1289.63,431.54)	(1294.26,434.66)	(1303.03,428.27)	(1293.77,430.83)
larceny	(2671.29,725.91)	(2679.68,711.38)	(2703.12,699.64)	(2707.34,708.84)	(2782.21,682.89)
auto-theft	(377.53,193.39)	(382.98,193.81)	(375.18,196.06)	(388.26,212.51)	(365.16,194.90)

表 7. 完整資料與各缺失比例下的主成份分析結果-主成份 1

Table 7. Principle component analysis of different missing data sets (principle 1)

principle 1	complete(0.588)	5%(0.610)	10%(0.609)	15%(0.605)	20%(0.619)
murder	0.300	0.294	0.308	0.615 [a]	0.104
rape	0.432 [a]	0.435 [a]	0.441 [a]	0.173	-0.008
robbery	0.397	0.401 [a]	0.400 [a]	-0.001	0.588 [a]
assault	0.397	0.395	0.406 [a]	0.335	-0.194

burglary	0.440 [a]	0.429	0.427 [a]	-0.185	-0.355
larceny	0.357	0.355	0.348	-0.393	-0.514 [b]
auto-theft	0.295	0.311	0.288	-0.539 [b]	0.465 [a]

a. The variable is greater than 0.4
b. The variable is smaller than -0.4

表 8. 完整資料與各缺失比例下的主成份分析結果-主成份 2

Table 8：Principle component analysis of different missing data sets (principle 2)

principle 2	complete(0.765)	5%(0.777)	10%(0.793)	15%(0.820)	20%(0.820)
murder	-0.629 [b]	0.664 [a]	-0.617 [b]	0.615 [a]	-0.595 [b]
rape	-0.169	0.115	-0.189	0.173	-0.178
robbery	0.042	-0.019	0.099	-0.001	0.029
assault	-0.344	0.351	-0.335	0.335	-0.350
burglary	0.203	-0.203	0.216	-0.185	0.172
larceny	0.402 [a]	-0.413 [b]	0.353	-0.393	0.416 [a]
auto-theft	0.502 [a]	-0.458 [b]	0.538 [a]	-0.539 [b]	0.537 [a]

a. The variable is greater than 0.4
b. The variable is smaller than -0.4

表 9. 完整資料與各缺失比例下的主成份分析結果-主成份 3

Table 9. Principle component analysis of different missing data sets (principle 3)

principle 3	complete(0.869)	5%(0.881)	10%(0.898)	15%(0.903)	20%(0.908)
murder	0.178	0.119	0.151	0.144	0.104
rape	-0.244	-0.208	-0.118	-0.124	-0.008
robbery	0.496 [a]	0.476 [a]	0.492 [a]	0.633 [a]	0.588 [a]
assault	-0.070	-0.087	-0.035	-0.210	-0.194
burglary	-0.210	-0.249	-0.281	-0.245	-0.355
larceny	-0.539 [b]	-0.521 [b]	-0.602 [b]	-0.508 [b]	-0.514 [b]
auto-theft	0.568 [a]	0.613 [a]	0.528 [a]	0.448 [a]	0.465 [a]

a. The variable is greater than 0.4
b. The variable is smaller than -0.4

第四節　因素分析(factor analysis)

Spearman(1904)提出因素分析，最早是用於心理學.，因心理學的領域中，常有無法直接測量或量化的因子，如智力、道德等，所以因素分析的目的即為經由可測量的變數來定義這些無法量化的因子。因素分析可視為主成份分析的延伸，目的為用較少、非直觀且隨機的因子來解釋收集到的資料。主要的用途是減少龐大的資料，並且可進行探索性的研究，找出潛在的變數，以供未來進一步的研究之用。主成份分析和因素分析兩者雖皆為資料縮減，但最主要的差異為，但主成份分析強調的解釋變異，找到指標。而因素分析則是解釋變數之間的關係，找出隱藏在背後的因素(Johnson，2007)。

本節採用資料同第三小節，不再贅述。分析結果列於表 8 及表 9，其中標記為 a 者代表因素荷重(factor loading)大於 0.7。

在完整資料中，可以看出因素 1 中，謀殺和暴力攻擊的因素荷重大於 0.7，故可將因素 1 視為人身攻擊類的變數，而因素 2 則是在偷竊上的因素荷重相當高，且入室竊盜的因素荷重雖然不到 0.7，但十分接近，故將因素 2 是為偷竊財物類的變數；而因素 3 則是每個因素荷重都沒有超過 0.7，無法提供一個合理的解釋。

表 11 為缺失 5%、10%、15%和 20%比例下進行因素分析所得之結果。由表中可以看出，在缺失 5%和 10%下，因素 1 和完整資料中因素荷重高的因子相同；而在缺失 15%和 20%下，因素荷重高的除了完整資料有的謀殺和暴力攻擊外，還多了一項強姦。在完整資料中，因素 1 中強姦的因素荷重為 0.676，和 0.7 雖然不算十分接近，但也相差不遠，所以可能造成當缺失比例增加時，該因子的因素荷重達到入選標準。即使如此，仍然可將因素 1 定義為人身攻擊類的變數，因為謀殺、強姦和暴力攻擊皆可是為人身攻擊的一種。

而在因素 2 的部份，除了缺失 5%之外，其他缺失比例下所得之因素 2 之結果都多一項竊盜。在完整資料的因素 2 中，竊盜的因素荷重為 0.689，和 0.7 接近，和因素 1 的強姦一樣，造成缺失比例增加時，荷重超過 0.7，進而達到入選標準。而不論是只有入室竊盜或是竊盜，因素二都可視為偷竊財物類的變數。

缺失 5%和 10%中的因素 3 和完整資料一樣，沒有因子的因素荷重超過 0.7；而缺失 15%和 20%時，則是強劫和偷車達到入選標準，但此兩因子無法歸成一類變數。整體來看，雖然在因素荷重的數字上沒有辦法完全一樣，但是所呈現出的結論卻是差異不大，皆為因素 1 為人身攻擊類的變數，因素 2 為偷竊財物類的變數，因素 3 無法有確切的分類。

表 10. 完整資料下的因素分析結果

Table 10. Factor analysis of complete data set

complete	factor 1	factor 2	factor 3	community
murder	0.815[a]	0.015	0.087	0.671
rape	0.676	0.520	0.231	0.782
robbery	0.472	0.242	0.625	0.672
assault	0.743[a]	0.310	0.208	0.697
burglary	0.390	0.689	0.466	0.843
larceny	0.119	0.817[a]	0.308	0.776
auto-theft	0.057	0.293	0.679	0.550
variance explained by each factor	2.065 (0.414)	1.659 (0.332)	1.268 (0.254)	total=4.991

a. The factor loading is greater than 0.7

表 11. 各缺失比例下之因素分析結果

Table 11. Factor analysis of different missing data sets

	missing 5%				missing 10%				missing 15%				missing 20%			
variable	factor 1	factor 2	factor 3	community	factor 1	factor 2	factor 3	community	factor 1	factor 2	factor 3	community	factor 1	factor 2	factor 3	community
murder	0.820 [a]	0.012	0.103	0.683	0.811 [a]	0.077	0.088	0.672	0.776 [a]	0.107	0.218	0.662	0.790 [a]	0.117	0.226	0.688
rape	0.665	0.512	0.250	0.767	0.636	0.542	0.239	0.756	0.771 [a]	0.468	0.214	0.860	0.702 [a]	0.533	0.339	0.892
robbery	0.432	0.236	0.641	0.654	0.493	0.206	0.649	0.707	0.509	0.197	0.750 [a]	0.859	0.554	0.221	0.756 [a]	0.927
assault	0.742 [a]	0.322	0.187	0.690	0.717 [a]	0.417	0.184	0.721	0.769 [a]	0.389	0.150	0.764	0.713 [a]	0.497	0.020	0.756
burglary	0.395	0.689	0.477	0.858	0.370	0.712 [a]	0.492	0.886	0.391	0.729 [a]	0.400	0.844	0.319	0.778 [a]	0.345	0.826
larceny	0.123	0.813 [a]	0.317	0.777	0.176	0.816 [a]	0.275	0.773	0.247	0.808 [a]	0.256	0.780	0.222	0.804 [a]	0.307	0.790
auto-theft	0.050	0.293	0.677	0.547	0.030	0.273	0.693	0.555	0.096	0.324	0.792 [a]	0.741	0.078	0.341	0.788 [a]	0.743
variance explained by each factor	2.026 (0.407)	1.643 (0.330)	1.306 (0.273)	total = 4.975	1.988 (0.392)	1.763 (0.348)	1.318 (0.260)	total = 5.070	2.271 (0.412)	1.711 (0.310)	1.530 (0.278)	total = 5.512	2.089 (0.372)	1.962 (0.349)	1.573 (0.280)	total = 5.623

a. The factor loading is greater than 0.7

第五節　鑑別分析(discrimination analysis)

鑑別分析為類別資料的分析工具，旨在將現有資料分類並找出分類的準則，以便新資料點出現可依各變數代入準則後歸於某一類中。為了使分類明確，若能使各資料群分得越開越好，以避免其間產生模糊地帶。Johnson(2007)指出在分類上很有可能會出現誤判的情形，並沒有完美分類方法存在，同時，在分類過程中也可能遇到各種問題增加分類的難度；好比說資訊不易得到或者成本高昂，如核磁共振等；也有可能要先將資料破壞才能得知屬於何類，如測量水果甜度時要先把水果切開等。

本文使用 SAS 9.1 User's Guild 中 Example 25.1：univariate density estimates and posterior probabilities 的資料為完整資料。現有一百五十朵鳶尾花，共分為三種品種，分別為 Setosa、Verisicolor 和 Virginica；欲就四變數建立一分類標準以區分三種不同品種之鳶尾花，其變數則為花萼長度、花萼寬度、花瓣長度和花瓣寬度四種。其分析結果列於表 10。

由表 12 可看出，原來每一品種的鳶尾花各有 50 朵，在分類後 Setosa 仍為 50 朵，而 Verisicolor 則為 49 朵，但其中僅 48 朵是分對的，一朵是因 Virginica 分錯而得，而 Verisicolor 分錯兩朵至 Virginica。因此算出，Setosa 的錯分率為 0 (0/50)，Verisicolor 為 0.04 (2/50)，而 Virginica 為 0.02 (1/50)；總錯分率為 0.02 (3/50)。

從表 13 及表 14 中看出，在各缺失比例下，Setosa 的錯分率皆為 0。但 Verisicolor 的錯分率就有變動，在缺失 5%和 20%時，Verisicolor 的分錯個數和完整資料相同，皆為兩朵，且皆錯分於 Virginica，錯分率為 0.04；而在 15%和 20%時，分錯的個數較少，為一朵，亦分錯於 Virginica，錯分率為 0.02。而 Virginica 則是隨著缺失比例的增加而增加，皆錯分為 Verisicolor，分別為一朵、兩朵、兩朵、四朵，而錯分率則各為 0.02、0.04、0.04、0.08。

整體錯分率則亦隨著缺失比例的增加而增加，分別為 0.02、0.02、0.033、0.04。雖然每一項目的錯分率不一定會隨著缺失比例的增加而增加，但是整體而言，錯分率的確如直觀判斷的跟隨著缺失比例的增加而增加。

表 12. 完整資料下之鑑別分析結果

Table 12. Discrimination analysis of complete data set

complete					
from species	Setosa	Versicolor	Virginica	total	error count estimates for species
Setosa	50	0	0	50	0
Versicolor	0	48	2	50	0.04
Virginica	0	1	49	50	0.02
total	50 (33.33)	49 (32.67)	51 (34)	150 (100)	0.02

表 13. 缺失比例為 5%和 10%下的鑑別分析結果

Table 13. Discrimination analysis from 5% and 10% missing data sets

	missing 5%					missing 10%				
from Species	Setosa	Versicolor	Virginica	total	error count estimates for species	Setosa	Versicolor	Virginica	total	error count estimates for species
Setosa	50	0	0	50	0	50	0	0	50	0
Versicolor	0	48	2	50	0.04	0	49	1	50	0.02
Virginica	0	1	49	50	0.02	0	2	48	50	0.04
total	50	49	51	150	0.02	50	51	49	150	0.02

表 14. 缺失比例為 15%和 20%下的鑑別分析結果

Table 14. Discrimination analysis from 15% and 20% missing data sets

	missing 15%					missing 20%				
from species	Setosa	Versicolor	Virginica	Total	error count estimates for species	Setosa	Versicolor	Virginica	total	error count estimates for species
Setosa	50	0	0	50	0	50	0	0	50	0
Versicolor	0	49	1	50	0.02	0	48	2	50	0.04
Virginica	0	4	46	50	0.08	0	4	46	50	0.08
total	50	53	47	150	0.03	50	52	48	150	0.04

第六節 多變量變異數分析(multivariate analysis of variance)

多變量變異數分析為比較多個母體均值的方法，但前題假設為所有母體的變異數矩陣Σ要相同，並且每個母體都是多元常態分布。其虛無假設為: 每個母體均值相同；統計模型為 $\boldsymbol{X}_{lj}=\boldsymbol{\mu}+\boldsymbol{\tau}_l+\boldsymbol{e}_{lj}$，其中 $j=1,2,...,n_l$、$l=1,2,...,g$，並且 e_{lj} 會服從 $N_p(0,\Sigma)$。在此模型下可將每個觀測值寫成 $\boldsymbol{x}_{lj}=\bar{\boldsymbol{x}}+(\bar{\boldsymbol{x}}_l-\bar{\boldsymbol{x}})+(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}}_l)$的形式，從而可得到如表 13 的 MANOVA Table (Johnson,2007)。

表 15. 多變量變異數分析表

Table15. MANOVA Table

source of variation	matrix of sum of squares and cross products	degrees of freedom
treatments (between)	$\boldsymbol{B}=\sum_{l=1}^{g}\boldsymbol{n}_l(\bar{\boldsymbol{x}}_l-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_l-\bar{\boldsymbol{x}})'$	g-1
residual (within)	$\boldsymbol{W}=\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}}_l)(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}}_l)'$	$\sum_{l=1}^{g}\boldsymbol{n}_l$ -g
total (corrected for the mean)	$\boldsymbol{B}+\boldsymbol{W}=\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}})(\boldsymbol{x}_{lj}-\bar{\boldsymbol{x}})'$	$\sum_{l=1}^{g}\boldsymbol{n}_l$ -1

從而得到 Wilks' Lambda(Λ^*)= $\frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|}$，服從 F 分配，自由度會隨著母體數(g)和變數個數(p)而有所不同，如當變數個數為 1，母體數大於等於 2 時，樣本分布為 $\left(\frac{\sum n_l-g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right)\sim F_{g-1,\sum n_l-g}$，若是母體數 g=2，變數個數 p 大於等於 1，則樣本分布為 $\left(\frac{\sum n_l-p-1}{p-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right)\sim F_{p,\sum n_l-p-1}$。

本文使用 SAS 9.1 User's Guild 中 Example 32.6：multivariate analysis of variance 的例子為例。測量英國四個不同窯中發現的古陶器中各化學成份是否相同，化學成份為鋁(Al)、鐵(Fe)、鎂(Mg)、鈣(Ca)和鈉(Na)。所得如表 16 所示。

表 16 中可看出，完整資料下所得之 P 值很小（<.0001），而不論缺失比例多少，所得之 P 值也都很小（<.0001），故採用 F 值做為判斷標準。直觀而言，當缺失比例越高所得之 F 值應越遠離完整資料，由表中可看出，在缺失比例為 15%時，是較其於三者更接近完整資料所得之 P 值，但由於完整資料得到相當顯著的結果，所以所得之結果仍相同，皆為拒絕虛無假設，即並非所有母體的均值都相同。

由於分析結果皆為顯著，表示並非所有母體的均值都相同，所以進一步的使用最小顯著差異測驗法(least significant difference test；LSD)進行多重比較，所得結果分別列於表 17，其中，以加註星號代表達到顯著差異。

首先，由表 17 的第一個部份可看出四個地區所出土陶器的鋁含量為 IslandThorns 和 AshleyRails 沒有顯著差異，並且 Llanederyn 和 Caldicot 之間亦沒有顯著差異，但是 IslandThorns 和 AshleyRailsy 則是明顯的大於 Llanederyn 和 Caldicot。並且，進一步的可由表 17 看出，在有達到顯著差異的組合中，計算所得的均值差並沒有受到缺失後插補的影響，在各缺失比例下數字都相當接近；然而，在沒有顯著差異的組合中，計算所得的均值差就受到缺失後插補的影響，數字變動較有達到顯著差異的組合來得大。

在表 17 的第二和第三部份中，四個地區所出土陶器的鐵含量為 Llanederyn 和 Caldicot 沒有顯著差異，而 IslandThorns 和 AshleyRails 也沒有顯著差異，但 Llanederyn 和 Caldicot 則是明顯的大於 IslandThorns 和 AshleyRails。而四個地區所出土陶器的鎂含量的結果和鐵含量相同。並且，由這兩張表中都可以發現，不論是有達到顯著的組合或是沒有達到顯著的組合其數字差異皆不大，都十分相近。

由表 17 的第四部份可看出四個地區所出土陶器的鈣含量為 Caldicot 顯著的大於其他三地；而 Llanederyn 又顯著的大於 IslandThorns 和 AshleyRails；IslandThorns 和 AshleyRails 兩者間則是沒有顯著差異。在表中可看到，當缺失比例小時，計算所得的均值差和完整資料所得的均值差是完全相同的，而在缺失比例大時才略有差距，但仍十分接近。推測可能原因為在隨機遺失的過程時，因子鈣含量在缺失比例小時沒有造成任何資料的遺失，或是遺失的個數很少，以致於不會影響到計算所得的均值差。

在表 17 的最後，可看出四個地區所出土陶器的鈉含量為 Llanederyn 顯著地大於其他三地，而 Caldicot、IslandThorns 和 AshleyRails 三者間則沒有顯著差異；從表中也可發現，不論有沒有顯著差異，各組合計算出的均值差皆於完整資料有不同，並不十分相近。綜合五張表的結果，可以推得，IslandThorns 和 AshleyRails 兩地出土陶器的各成份都沒有顯著差異，表示兩地出土的陶器成份相近，和 Llanederyn 和 Caldicot 的成份差異較大；而 Llanederyn 和 Caldicot 兩地出土的陶器在成份鋁、鐵和鎂三者上沒有顯著差異，但在鈣和鈉上則有明顯的差異，兩地的成份雖然相近，但相似的程度不若 IslandThorns 和 AshleyRails 兩地相似程度之高。

表 16. 完整資料與各缺失比例下多變量變異數分析之結果

Table 16. MANOVA analysis from complete data set

	value	F value	Pr > F
complete	0.01230091	13.09	<.0001
missing 5%	0.02205943	9.96	<.0001
missing 10%	0.02332831	9.69	<.0001
missing 15%	0.02010070	10.41	<.0001
missing 20%	0.02633967	9.13	<.0001

表 17. 完整資料與各缺失比例下多重比較之結果

Table 17. Multiple comparison of different missing data sets

Al	complete	5%	10%	15%	20%
IslandThorns - AshleyRails	0.860	1.475	1.163	1.198	1.274
IslandThorns - Llanederyn	5.616*	5.880*	5.770*	5.752*	5.606*
IslandThorns - Caldicot	6.480*	6.480*	7.104*	6.797*	6.543*
AshleyRails - Llanederyn	4.756*	4.405*	4.607*	4.554*	4.332*
AshleyRails - Caldicot	5.620*	5.006*	5.941*	5.599*	5.269*
Llanederyn - Caldicot	0.864	0.600	1.334	1.045	0.937

表 17. 完整資料與各缺失比例下多重比較之結果(續)

Table 17. Multiple comparison of different missing data sets (continued)

Fe	complete	5%	10%	15%	20%
Llanederyn - Caldicot	0.957	0.787	0.812	0.823	0.968
Llanederyn - IslandThorns	4.660*	4.490*	4.515*	4.795*	4.825*
Llanederyn - AshleyRails	4.860*	4.690*	4.715*	4.986*	5.049*
Caldicot - IslandThorns	3.703*	3.703*	3.703*	3.971*	3.857*
Caldicot - AshleyRails	3.903*	3.903*	3.903*	4.162*	4.081*
IslandThorns - AshleyRails	0.200	0.200	0.200	0.191	0.224

表 17. 完整資料與各缺失比例下多重比較之結果(續)

Table 17. Multiple comparison of different missing data sets (continued)

Mg	complete	5%	10%	15%	20%
Llanederyn - Caldicot	0.971	0.971	1.157	1.104	0.979
Llanederyn - IslandThorns	4.152*	4.152*	4.338*	4.149*	4.160*
Llanederyn - AshleyRails	4.220*	4.220*	4.473*	4.285*	4.304*
Caldicot - IslandThorns	3.181*	3.181*	3.181*	3.045*	3.181*
Caldicot - AshleyRails	3.249*	3.249*	3.316*	3.181*	3.325*
IslandThorns - AshleyRails	0.068	0.068	0.135	0.136	0.144

表 17. 完整資料與各缺失比例下多重比較之結果(續)

Table 17. Multiple comparison of different missing data sets (continued)

Ca	complete	5%	10%	15%	20%
Caldicot - Llanederyn	0.093*	0.093*	0.093*	0.099*	0.094*
Caldicot - AshleyRails	0.243*	0.243*	0.243*	0.243*	0.242*
Caldicot - IslandThorns	0.269*	0.269*	0.269*	0.269*	0.258*
Llanederyn - AshleyRails	0.150*	0.150*	0.150*	0.144*	0.147*
Llanederyn - IslandThorns	0.176*	0.176*	0.176*	0.170*	0.164*
AshleyRails - IslandThorns	0.026	0.026	0.026	0.026	0.017

表 17. 完整資料與各缺失比例下多重比較之結果(續)

Table 17. Multiple comparison of different missing data sets (continued)

Na	complete	5%	10%	15%	20%
Llanederyn - IslandThorns	0.197*	0.209*	0.198*	0.193*	0.190*
Llanederyn - Caldicot	0.201*	0.203*	0.214*	0.191*	0.195*
Llanederyn - AshleyRails	0.203*	0.205*	0.216*	0.193*	0.204*
IslandThorns - Caldicot	0.004	-0.007	0.016	-0.003	0.005
IslandThorns - AshleyRails	0.006	-0.005	0.018	-0.001	0.014
Caldicot - AshleyRails	0.002	0.002	0.002	0.002	0.009

第七節　典型相關分析(canonical analysis)

Hotelling(1936)提出典型相關分析，用以研究兩組變數其線性組合結果是否有相關。與一般的相關係數不同的地方是，一般相關係數計算一對一的相關程度，典型相關係數是計算多對多的相關程度。分別找出 p 個 X 變數與 q 個 Y 變數的線性組合，使 p 個 X 變項的線性組合與 q 個 Y 變項的線性組合兩者間的相關係數達到最大值，此一相關係數即為典型相關係數(canonical correlation)，藉著典型相關係數即可判斷兩組變數是否有關連。在進行典型相關分析時，在找出第一對相關程度最大的線性組合後，還可以找出與第一對線性組合沒有相關的第二對線性組合，這第二對線性組合相關程度次高的線性組合，可依此繼續找出第 m 對線性組合。

假設有兩組隨機向量 **X**、**Y**，其平均和共變數矩陣分別為 $E(\boldsymbol{X})=\boldsymbol{\mu}_x$，$E(\boldsymbol{Y})=\boldsymbol{\mu}_y$，$Cov(\boldsymbol{Y})=\boldsymbol{\Sigma}_y$，$Cov(\boldsymbol{X})=\boldsymbol{\Sigma}_x$；而兩變數的共變數矩陣則為 $Cov(\boldsymbol{X},\boldsymbol{Y})=\boldsymbol{\Sigma}_{xy}$。將隨機向量經過線性組合後分別得 $U=\boldsymbol{a}'\boldsymbol{X}$ 和 $V=\boldsymbol{b}'\boldsymbol{Y}$，稱之為典型變數(canonical variables)；而後可以計算出兩者的變異數和共變數為 $Var(U)=\boldsymbol{a}'\boldsymbol{\Sigma}_x\boldsymbol{a}$，$Var(V)=\boldsymbol{b}'\boldsymbol{\Sigma}_y\boldsymbol{b}$，$Cov(U,V)=\boldsymbol{a}'\boldsymbol{\Sigma}_{xy}\boldsymbol{b}$，從而計算出典型相關係數 $Corr(U,V)=\dfrac{\boldsymbol{a}'\boldsymbol{\Sigma}_{xy}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_x\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_y\boldsymbol{b}}}$ (Johnson,2007)。

使用 SAS 9.1 User's Guild 中 Example 20.1: canonical correlation analysis of fitness club data 的資料為完整資料。測量 20 位中年人的身體情況和體能狀況。變數為：體重(weight)、腰圍 waist)、脈搏數(pulse)、引體向上(chins)、仰臥起坐(situps)、跳遠(jumps)。將這些變數分為兩類，一類為身體狀況（physiological measurements）：體重、腰圍、脈搏數；另一類則為體能狀況（exercises）；引體向上、仰臥起坐、跳遠。典型相關分析則是要判斷此兩類中變數線性組合是否相關，將屬於身體情況三個變數的線性組合令為 U，而屬於體能狀況的三個變數的線性組合令為 V，以計算處此二典型變數最大的典型相關係數。

由表 18 來看，在完整資料下找到最大典型相關係數約為 0.80，在顯著水準為 0.05 下無法拒絕需無假設，所以沒有足夠的證據說明兩典型變數間有相關性存在。而缺失 5%、10%和 15%下，所得到的最大典型相關係數分別為 0.88、0.88 和 0.87，三者相當接近，但與完整資料相比，有些許差異。再由 P 值看出當缺失 5%、10%和 15%下，P 值僅為 0.0023、0.0015 和 0.0055，不但和完整資料所得結果差異大，並且已經小於顯著水準 0.05，代表有足夠的證據說明兩個典型變數間有相關性存在。

分別單獨看典型變數 U 和 V 中各變數的情形。在完整資料下，體重和腰圍分別是典型變數 U 的負因子（-0.7754）和正因子（1.5793），而脈搏數接近 0（-0.0591），故對 U 幾乎沒有影響。在缺失 5%、10%和 15%下，亦得到相同結果，但其數字略有差異，其中又以缺失 15%時所得之數字差異最大(分別為-0.5423、1.4239 和-0.0227)。

而典型變數 V，不論是完整資料或是有缺失的資料，方向和數字都差異不大。引體向上的數值較小，但距 0 仍有些許距離，只能判斷該因子對 V 的影響較小，不能算沒有影響。而仰臥起坐則是 V 的負因子，其數值皆靠近-1；而跳遠則視為 V 的正因子，但其數字變動範圍較大，從 0.716 到 1.27。

表 18. 完整資料與各缺失比例下典型相關分析結果

Table18. Canonical correlation analysis of different missing data sets

	canonical correlation	Wilks' Lambda	U=(Physiological)			V=(Exercises)		
		Pr > F	weight	waist	pulse	chins	situps	jumps
complete	0.795608	0.0635	-0.7754	1.5793	-0.0591	-0.3495	-1.0540	0.7164
missing 5%	0.884887	0.0023	-0.8481	1.5544	-0.0524	-0.3395	-1.0991	1.1547
missing 10%	0.886916	0.0015	-0.8574	1.5872	-0.0176	-0.3981	-1.0803	1.2753
missing 15%	0.873771	0.0055	-0.5423	1.4239	-0.0227	-0.5344	-1.0806	0.9940

第八節 共變數結構(covariance structure)

SAS 在進行多重插補時皆假設母體為多元常態分布，現在欲比較不同共變數結構下，使用多重插補法會不會造成共變數結構的改變。根據四種不同型態的共變數結構，產生一筆各數為 100 新資料，再經過隨機缺失 10%和 20%後，比較完整資料和缺失資料的共變數是否相同。此四種共變數結構分別為任意的共變數結構(unstructured covariance structure)、混合對稱的共變數結構(compound symmetry covariance structure)、第一級自我迴歸的共變數結構(first-order autoregressive covariance structure)和 Toeplitz 氏的共變數結構(Toeplitz covariance structure)。

比較兩共變數是否相同採用之統計量如下：先計算 $L = v(\ln|\Sigma_0| - \ln|S| + trS\Sigma_0^{-1} - p)$，其中 v 為 S 的自由度，在這裡為 100-1=99，Σ_0 為假定的共變數矩陣，S 為要比較的矩陣，p 則為變數數目，在這裡為 5。當觀測值很大時，L 會趨近自由度為 $\frac{1}{2}p(p+1)=15$ 的卡方分布。從而計算出 $L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}$；當 L1 大於 $\chi^2_{0.05,15}$ 時，即拒絕虛無假設，代表兩共變數矩陣不相同（Morrison,1990）。

1. 任意的共變數結構為沿軌跡(trace)對稱(SAS，2002)，如下

$$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

根據其結構令一共變數為表 19 左側之共變數矩陣，右側則為完整資料下所計算出的共變數矩陣，所得之卡方值約為 0.4，符合要求。而表 20 左右兩矩陣分別代表缺失 10%和 20%時所得之共變數矩陣。雖得到的矩陣數字和完整資料不完全相同，但是計算所得之卡方值皆大於 0.05，因此判斷在共變數結構為任意的共變數結構，缺失後插補並沒有影響共變數結構。

表 19. 母體共變數及完整資料所得之共變數估計值 (任意的共變數結構)

Table 19. Covariance parameters and estimates from complete data set (unstructured covariance structure)

covariance structure of original setting					complete				
					L1[a]：13.324509 Prob($\chi^2_{df=15}$>L1)=0.4227514				
9	2	3	1	3	7.787	1.829	2.199	1.708	3.226
2	5	2	1	2	1.829	3.542	2.323	1.248	1.147
3	2	7	2	1	2.199	2.323	7.360	2.431	0.587
1	1	2	8	3	1.708	1.248	2.431	7.818	2.963
3	2	1	3	6	3.226	1.147	0.587	2.963	5.741

a：$L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}\times$

表 20. 缺失 10%和 15%下所得之共變數估計值 (任意的共變數結構)

Table 20. Covariance estimates from 10% and 20% missing data sets (unstructured covariance structure)

missing 10%					missing 20%				
L1[a]：12.583769 Prob($\chi^2_{df=15}$>L1)=0.3655873					L1[a]：14.694917 Prob($\chi^2_{df=15}$>L1)=0.5263921				
9.401	2.291	3.064	2.370	2.581	9.861	1.968	2.977	-0.374	4.031
2.291	5.276	2.526	2.435	2.459	1.968	5.000	1.899	1.897	2.794
3.064	2.526	6.135	2.798	1.029	2.977	1.899	7.316	2.567	1.893
2.370	2.435	2.798	8.408	3.666	-0.374	1.897	2.567	9.845	3.858
2.581	2.459	1.029	3.666	6.001	4.031	2.794	1.893	3.858	7.010

a：$L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}$

2. 混合對稱的共變數結構為沿軌跡對稱，且除了軌跡外，其餘數字要相同(SAS, 2002)。

$$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$$

根據其結構令一共變數為表 21 左側之共變數矩陣，右側則為完整資料下所計算出的共變數矩陣，所得之卡方值約為 0.1，符合要求。而表 22 左右兩矩陣分別代表缺失 10%和 20%時所得之共變數矩陣。得到的矩陣數字和完整資料不完全相同但十分接近，且計算所得之卡方值皆大於 0.05，因此判斷在共變數結構為混合對稱的共變數結構下，缺失後插補並沒有影響共變數結構。

表 21. 母體共變數及完整資料所得之共變數估計值
(混合對稱的共變數結構)

Table 21. Covariance parameters and estimates from complete data set
(compound symmetry covariance structure)

covariance structure of original setting					complete				
					L1[a]：9.9523715 Prob($\chi^2_{df=15}$>L1)=0.1772749				
5	2	2	2	2	4.984	2.151	1.338	2.014	1.643
2	5	2	2	2	2.151	4.924	1.779	1.385	1.265
2	2	5	2	2	1.338	1.779	4.594	1.593	1.825
2	2	2	5	2	2.014	1.385	1.593	4.967	1.593
2	2	2	2	5	1.643	1.265	1.825	1.593	4.087

a：$L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}$

表 22. 缺失 10%和 15%下所得之共變數估計值 (混合對稱的共變數結構)

Table 22. Covariance estimates from 10% and 20% missing data sets (compound symmetry covariance structure)

missing 10%					missing 20%				
L1[a]：18.409009 Prob($\chi^2_{df=15}$>L1)=0.7582265					L1[a]：14.120201 Prob($\chi^2_{df=15}$>L1)=0.4835669				
4.789	1.732	1.420	0.908	1.419	4.758	1.694	2.322	2.558	1.106
1.732	4.438	1.504	1.550	1.413	1.694	4.007	1.919	1.698	1.168
1.420	1.504	4.308	1.517	1.901	2.322	1.919	5.393	2.358	1.644
0.908	1.550	1.517	3.043	1.212	2.558	1.698	2.358	4.660	1.550
1.419	1.413	1.901	1.212	4.870	1.106	1.168	1.644	1.550	4.452

a：$L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}$

3. 第一級自我迴歸的共變數結構為提出σ^2後，軌跡為 1 而和右側即下方之數值相差ρ倍(SAS, 2002)。

$$\sigma^2\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

根據其結構令一共變數為表 23 時左側之共變數矩陣，右側則為完整資料下所計算出的共變數矩陣，所得之卡方值約為 0.8，表是兩共變數矩陣相同。而表 24 左右兩矩陣分別代表缺失 10%和 20%時所得之共變數矩陣。得到的矩陣數字和完整資料有些許差異，計算所得之卡方值皆分別為 0.767、0.276，因此判斷在共變數結構為第一級自我迴歸的共變數結構下，缺失後插補並沒有影響共變數結構。

表 23. 母體共變數及完整資料所得之共變數估計值

(第一級自我迴歸的共變數結構)

Table 23. Covariance parameters and estimates from complete data set

(first-order autoregressive covariance structure)

covariance structure of original setting					complete				
					L1[a]：19.596399 Prob($\chi^2_{df=15}$>L1)=0.8120321				
1	1/2	1/4	1/8	1/16	1.314	0.503	0.157	-0.010	0.128
1/2	1	1/2	1/4	1/8	0.503	0.865	0.546	0.238	0.140
1/4	1/2	1	1/2	1/4	0.157	0.546	1.019	0.491	0.346
1/8	1/4	1/2	1	1/2	-0.010	0.238	0.491	1.044	0.481
1/16	1/8	1/4	1/2	1	0.128	0.140	0.346	0.481	0.956

a：$L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}$

表 24. 缺失 10%和 15%下所得之共變數估計值

(第一級自我迴歸的共變數結構)

Table 24. Covariance estimates from 10% and 20% missing data sets

(first-order autoregressive covariance structure)

missing 10%					missing 20%				
L1[a]：18.585392 Prob($\chi^2_{df=15}$>L1)=0.7668598					L1[a]：11.396217 Prob($\chi^2_{df=15}$>L1)=0.275965				
1.253	0.737	0.378	0.311	0.023	1.074	0.527	0.238	0.137	0.037
0.737	1.168	0.596	0.321	0.008	0.527	1.112	0.473	0.278	0.125
0.378	0.596	1.025	0.639	0.302	0.238	0.473	0.895	0.456	0.267
0.311	0.321	0.639	1.017	0.458	0.137	0.278	0.456	0.855	0.456
0.023	0.008	0.302	0.458	0.785	0.037	0.125	0.267	0.456	0.750

a：$L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}$

4. Toeplitz 氏的共變數結構為沿軌跡對稱，且軌跡上各值要相同(SAS, 2002)。

$$\begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

根據其結構令一共變數為表 25 左側之共變數矩陣，右側則為完整資料下所計算出的共變數矩陣，所得之卡方值約為 0.4，表是兩共變數矩陣相同。而表 26 左右兩矩陣分別代表缺失 10%和 20%時所得之共變數矩陣數值和完整資料相近，但兩者的結果卻不相同。在缺失比例為 10%下，其卡方值為 0.6，代表矩陣仍然後原始設定之共變數矩陣相同。而在缺失 20%時，數字雖接近完整資料，但計算所得之卡方值卻只有 0.004，代表和原始設定之共變數矩陣不同。因此在 Toeplitz 氏的共變數結構下，缺失比例達 20%時，多重插補法可能會影響原始資料的共變數結構。

表 25. 母體共變數及完整資料所得之共變數估計值

(Toeplitz 氏的共變數結構)

Table 25. Covariance parameters and estimates from complete data set (Toeplitz)

covariance structure of original setting					complete				
					L1[a]：12.648332 Prob($\chi^2_{df=15}$>L1)=0.3705599				
6	1	2	3	4	5.719	1.418	1.906	2.244	4.338
1	6	1	2	3	1.418	7.191	0.242	2.477	4.249
2	1	6	1	2	1.906	0.242	5.678	-0.030	1.798
3	2	1	6	1	2.244	2.477	-0.030	5.425	0.811
4	3	2	1	6	4.338	4.249	1.798	0.811	7.150

a：$L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}$

表 26. 缺失 10%和 15%下所得之共變數估計值 (Toeplitz 氏的共變數結構)

Table26. Covariance estimates from 10% and 20% missing data sets (Toeplitz)

missing 10%					missing 20%				
L1[a]：15.860604 Prob($\chi^2_{df=15}$>L1)=0.6086239					L1[a]：4.4931454 Prob($\chi^2_{df=15}$>L1)=0.0043825				
5.649	0.895	2.767	2.284	4.034	5.945	0.697	2.201	3.576	3.759
0.895	6.944	1.926	2.321	3.310	0.697	5.994	0.697	1.949	2.956
2.767	1.926	6.466	0.914	2.907	2.201	0.697	5.794	0.991	2.081
2.284	2.321	0.914	4.796	1.421	3.576	1.949	0.991	6.737	1.453
4.034	3.310	2.907	1.421	6.450	3.759	2.956	2.081	1.453	5.907

a：$L1=\left\{1-\frac{1}{6(N-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}$

第四章　結果與討論

本文比較了多種不同統計分析方法下使用多重插補法後結果的異同。模擬結果的第一部份為可估計母體參數的統計分析方法，分為迴歸分析和羅吉斯迴歸兩部份；第二部份為不估算母體參數的分析方法，為主成份分析、因素分析、鑑別分析、多變量分析和典型相關分析五種；第三部份則是四種共變數結構在缺失後插補的比較。

第一節　估計母體參數

迴歸分析的結果發現，當缺失比例越高時，P 值會隨之變動，且離完整資料越來越遠。在缺失比例較低(5%)的情況下，不但篩選出的變數相同，並且除了靜止時脈搏之外，其他計算所得 P 值的數字差異也不大；在缺失比例為 10%開始和完整資料就出現差距，變動幅度明顯的增加許多，但是方法 A 和 B 在缺失比例 10%時篩選出的變數相同，無優劣之分，但是靜止時脈搏數和最大脈搏數兩變數所得之 P 值有較大的差距，分別為 0.3287 和 0.2430；在缺失比例為 15%時，可以看出方法 B 較優於方法 A，因為截距項在完整資料下是顯著的，可是方法 A 並沒有將其篩選出來，同時也可看出方法 A 和 B 的 P 值差異變大；而在 20%時，方法 A 少篩選出跑步時脈搏數和最大脈搏數，方法 B 則是多篩選出靜止時脈搏數少篩選出跑步時脈搏數。

經由這個比較，可得到如下的結論，缺失比例低時，完整資料、方法 A 和方法 B 三者篩選出來的變數不會有太大的差異；但當缺失比例越來越高時，無論是方法 A 或是方法 B 都無法得到和母體相近的結果，並且方法 A 和方法 B 兩者間的差距也變大，同時，可以發現，在方法 B 下得到顯著的變項較多(在缺失比例為 15%和 20%時)，因此，若是在實驗初期或是進行試驗性研究（pilot study），不想錯過任何一個因子時，可使用方法 B 進行分析，以免遺漏掉重要的變項，而在實驗的最後階段可使用方法 A，以便選出最具有代表性的因子。

從羅吉斯迴歸分析的結果可以看出在缺失比例為 5%和 10%時，和完整資料分析所得之 P 值數字差異較小，而 15%和 20%時，不單和完整資料有明顯

差異，方法 A 和 B 之間也有明顯的差異。由於在羅吉斯迴歸中採用的例子在完整資料下沒有任何一個變項達到顯著，因此由各缺失比例下，方法 A 和 B 所得之 P 值與完整資料所得之 P 值相減，數字越小代表越相近，比較方法 A 和 B 與完整資料相近的變數較多，就判定該方法較好。在缺失比例為 5%時，發現方法 A 除了 li 和 temp 兩者之外，其餘變數皆小於方法 B，因此判斷方法 A 較好；在缺失比例為 10%時，以同樣的方法做判斷，則是方法 B 較好；而在 15%和 20%中，兩者結果不相上下。綜合以上結果來看，無法判斷方法 A 和 B 的優劣，但是可以發現使用羅吉斯迴歸分析時，若缺失比例很大，採用多重插補法是較不恰當的。

在可估計母體參數的兩種統計分析結果中，發現一開始為不顯著的因子，其後改變結果得到顯著的情況很少，如羅吉斯迴歸中的各變數；而非常顯著（$p<0.0001$）的因子之後也少有變為不顯著的情況，如迴歸分析中的跑步時間；但若是靠近臨界值的變數，其變動範圍便相當大，有時達到顯著有時卻無法被篩選出來，如迴歸分析中的年齡和最大脈搏數。

第二節　非估計母體參數

主成份分析中，不論缺失比例為何，其插補後所得的平均和完整資料計算所得的平均十分相近。而七個變數中，和完整資料平均最相近的不一定是缺失比例較小的。 如謀殺對應的是缺失比例為 10%，其和完整資料的差距為 0.01(單位：百萬人分之一)；強姦對應的是 15%，和完整資料的差距為 0.08，搶劫、 暴力攻擊、入室竊盜、偷竊以及偷車等變數分別對應到的缺失比例為 20%、10%、20%、5%和 10%，和完整資料的差距分別為 0.17、0.22、1.87、8.39 和 2.35。發現在此例中，和完整資料最接近的變數多落在缺失 10%時，共有三變數，謀殺、暴力攻擊和偷車；其次是 20%，為搶劫和入室竊盜。

若按各變數下四種缺失比例插補所得的平均由小到大排序，從 1 到 4，1 為最接近者。可以得到如表 27 所呈現的結果。而加總值越小表示與完整資料越接近。因此可得一結論，不一定是缺失比例低時，結果才會和完整資料相近，以此例而言，是在 10%時和完整資料最相近。若不論對新變數的貢獻是

正向或負向，僅由數字來判斷該因子對新變數是否有達到顯著(大於 0.4)，則缺失 5%是和完整資料最相近的，只有在主成份 1 時多一因子--強劫達到顯著。倘若將正負向共獻一併納入考慮時，則發現缺失比例 10%和完整資料所得結果最接近，而這樣的結果和排序後的結果是一致的。而缺失比例為 15%和 20%時，雖然主成份 1 和完整資料差異相當大，但是在主成份 2 及主成份 3 時，卻又和完整資料十分接近。

因素分析和主成份分析使用同一筆資料進行分析，在缺失比例 5%時，所得之結果和完整資料一樣，荷重大於 0.7 的為因素 1 下的謀殺和暴力攻擊，和因素二中的偷竊。在缺失比例 10%時，除了完整資料所得之因子外，因素荷重大於 0.7 的還有因素 2 中的入室竊盜。而 15%和 20%下，因素荷重大於 0.7 者變多，但是在完整資料中已被選出的因子(荷重大於 0.7)並沒有因為缺失比例的增加而降低因素荷重至被淘汰。

由於主成份分析和因素分析的資料相同，將此兩方法作比較，發現在因素分析中插補後所得的結果和完整資料下所得的結果較為相近，雖然在缺失 15%和 20%時，達到入選水準的因子增加不少，但其中有些因子的因素荷重在完整資料時所得到的數值十分接近 0.7，如強姦(0.676)和入室竊盜(0.689)這兩項，在缺失插補後可能會造成些許變動，致使這些因子達到入選標準。除此之外，不論在何種缺失比例下，對於三個因素的判斷並沒有改變，因素 1 為人身攻擊的變數，因素 2 為偷竊財物的變數，因素 3 則是無法有確切的分類。相較於因素分析，主成分分析插補後所得結果和完整資料所得結果的差異略大，特別是當缺失比例增加到 15%以上時，差異十分明顯。

鑑別分析中，在完整資料下完全沒有被錯分的 Setosa 不論後來缺失比例為何，都不會被錯分，表示 Setosa 這一品種的花是相當獨特，並且建立出的分類標準是可以完全的將所有 Setosa 和其他的品種的花分離。但在完整資料下 Versicolor 和 Virginica 就已經錯分至彼此的品種中，由此可知，此兩品種的花較相似，以致於分類標準無法完美的將此兩品種分開。

因此，當缺失比例增加時，Versicolor 和 Virginica 這兩種本來就比較相似的品種中，錯分的個數亦隨之增加，並且稍有不同。如 Versicolor 被分錯的個數有多(2 朵)有少(1 朵)，但 Virginica 則是隨著缺失比例的增加而錯分的

個數也增加(由 1 朵到 4 朵),可能是因為這些花朵一開始就落在分類的模糊地帶上，因此隨著缺失比例的增加而產生不同結果。而總體錯分率是隨缺失比例增加而穩定增加(由 0.02 到 0.04)，不過總體錯分率很小。由此看來，多重插補法在鑑別分析中不論缺失比例皆可以得到不錯的效果，但是也有可能是因為這三個品種的花本身差異就相當明顯，因此建立的分類準則也就不會產生太大的變動，花朵的分類也就不會有太大的差別。

在多變量變異數分析中，完整資料下的分析結果為非常顯著($p<0.0001$)，因此之後使完整資料缺失插補仍然得到非常顯著的結果，即古陶器中五種不同的化學成份含量不完全相同。雖然結果相同，但是仍可以用 F 值判斷各缺失比例和母體的差異為何，缺失 5%、10%、15%和 20%的情況下，F 值和完整資料的差距分別為 3.13、3.4、2.68 和 3.96。發現在缺失比例為 15%時，F 值和完整資料最相近，和直觀上缺失比例越低和完整資料會越相近的看法不同。

進一步進行多重比較時發現，不論在何種缺失比例下有達到顯著水準的組合都相同，沒有改變。四地區出土陶器的成份鋁、鐵、鎂、鈣或者是鈉含量的比較下，所得結果並沒有因為缺失比例的增加而有所不同，在鋁含量的部份，得到的結果為 IslandThorns ~ AshleyRails > Llanederyn ~ Caldicot，其中符號"~"代表沒有顯著差異。在鐵含量及鎂含量所得結果皆為 Llanederyn ~ Caldicot > IslandThorns ~ AshleyRails。而鈣含量的結果為 Caldicot > Llanederyn > IslandThorns ~ AshleyRails。鈉含量所得的結果為 Llanederyn > Caldicot ~ IslandThorns ~ AshleyRails。而在各成份下計算所得之均值差，也沒有因為缺失比例的增加而有太大的差異，其中，受到缺失影響最大的為鈉含量的因子，鈣含量則是最沒有受到缺失影響，但不論如何，所得之結果和完整資料都十分相近。

典型相關分析的結果中可以發現，三種缺失比例下的典型相關係數十分接近，但是與完整資料的典型相關係數(0.796)卻有差距存在。從 P 值來看，可以發現，完整資料是無法拒絕虛無假設的，即沒有足夠的證據證明兩典型變數間有相關性存在，不過雖然無法拒絕虛無假設，但是 P 值卻是十分靠近臨界值的($p=0.0635$)，所以在缺失插補後使得各缺失比例都有達到顯著水準。

由典型變數 U 和 V 來看，U 下的三個變數體重、腰圍和脈搏數中，和完整資料最接近的是腰圍；而體重則是除了缺失 15%外，其餘二者和完整資料相近；脈搏數則是只有缺失比例 5%的時候和完整資料相近，其餘二者和完整資料的差別較明顯。V 下的三個變數引體向上、仰臥起坐和跳遠，和完整資料最接近的是仰臥起坐；其次是引體向上，該變數在缺失比例為 15%時，和完整資料的差異明顯，其餘二者和完整資料則是十分接近；而跳遠則是不論在何種比例下，和完整資料的差距都相當大。

理論上，缺失值比例越大和完整資料差異也越大，但實際上則否，可能原因為插補時僅使用一次起始值去計算，並沒有重複多次插補求得較穩定的結果，因而造成差異的變動並非隨著缺失比例增加而增加，而是有時增加、有時減少。

在非估計母體參數的統計方法中，發現缺失插補後的分析結果最接近的為因素分析，其次為鑑別分析、多變量分析和典型相關三者，主成份分析和完整資料的差異最為明顯。

表 27. 主成份分析中插補後均值與完整資料均值差之排序

Table 27. The order of the difference between the mean of different missing data sets and of complete data set from principle component analysis

	5%	10%	15%	20%
murder	2	1	4	3
rape	4	2	1	3
robbery	3	4	2	1
assault	4	1	2	3
burglary	2	3	4	1
larceny	1	2	3	4
auto theft	2	1	3	4
total	18	14	19	19

第三節　共變數結構

在四種不同的共變數結構任意的共變數結構、混合對稱的共變數結構 、第一級自我迴歸的共變數結構和 Toeplitz 氏的共變數結構下，比較缺失比例為 10%和 20%下，會不會因為缺失插補後而改變其共變數結構。比較分析結果為在任意的共變數結構可以發現，10%和 20%下所得的共變數矩陣在數值部分較相似，但和完整資料下的矩陣數值差異很大，不過由統計值 L1 計算出的 P 值卻很大，分別為 0.3656 和 0.5264，不能拒絕虛無假設，表示這兩個矩陣都符合原來設定的母體結構，並沒有因為缺失後插補而造成共變數矩陣型態的改變。

以同樣的方式比較混合對稱的共變數結構的結果，所得到兩矩陣數值部分不但接近,和完整資料矩陣的數值部分也相近,再由 L1 所得到的 P 值來看，兩者分別為 0.7582 和 0.4836，不能拒絕虛無假設，表示這兩個矩陣都符合原來設定的母體結構，並沒有因為缺失後插補而造成共變數矩陣型態的改變。

在第一級自我迴歸的共變數結構中，則是缺失 20%時所得到的數值較缺失 10%時更接近完整資料，一樣的由 L1 計算所得到的 P 值來看，兩者分別為 0.769 和 0.276，不能拒絕虛無假設，在第一級自我迴歸的共變數結構中，不會因為缺失後插補造成共變數結構的改變。

最後一個共變數結構為 Toeplitz 氏的共變數結構，雖然兩者所得到的矩陣和完整資料較為相似，但由 L1 計算所得到的結果卻不是如此。缺失 10%和缺失 20%所得到的 P 值分別為 0.6083 和 0.004。代表缺失比例為 10%時，插捕後計算得到的矩陣結構並沒有改變，仍然是 Toeplitz 氏的共變數結構；但是在缺失比例為 20%時，P 值小於顯著水準 0.05，代表缺失比例為 20%時，插補後所得到的共變數結構不同於完整資料，意味 Toeplitz 氏的共變數結構可能會因缺失後插補改變共變數結構。

綜合來看共變數結構若為任意的共變數結構、混合對稱的共變數結構 和第一級自我迴歸的共變數結構三者不會因為缺失後插補而改變，但是在 Toeplitz 氏的共變數結構時便有可能因為缺失比例過高而造成共變數結構的改變。而與完整資料比較後也可發現，CS 和 AR 兩者無分軒輊，和完整資料

相似性頗高；其次為任意的共變數結構，有些微的差異；最後是 Toeplitz 氏的共變數結構，缺失比例高時，共變數結構會改變。

第四節 結論及後續研究建議

由上述的結果中可發現，在原始的完整資料分析下，得到非常顯著的因子在缺失後插補仍然會維持非常顯著，如多變量變異數分析中的例子；而非常不顯著的因子在缺失後插補可維持其不顯著的特性，如羅吉斯迴歸的例子；而最容易受到缺失影響的為靠近臨界值附近的因子，插補完成後的結果常常會有所變動，有時達到顯著，有時卻不顯著，如迴歸分析例子中，年齡的變項，就會造成有時顯著，有時不然的結果，。

理論上，缺失值比例越大和完整資料差異也越大，但實際上則不一定，如主成分分析的例子，當缺失比例為 10%時，所得到的平均數和完整資料最為接近，其次才是 5%，而 15%和 20%兩者則是並列第三名，由此可知，不一定缺失比例越高，分析結果就越不可靠。

另外，多重插補法在估計效率上要判斷其效率有一定的困難，難以計算在何種缺失比例下使用多重插補法可以穩當無誤，由文中也可以發現，不一定缺失比例越高，計算出的結果就會與完整資料差異越大。還是應該要先確認前提假設是否完備，多使用不同的起始值最計算，重複施行多重插補法分析，或者將多重插補法的套數增加，以求得到穩定可靠的結果。

在 SAS 中多重插補法前提為資料形態需為多元常態分布，但事實上，多重插補法的應用範圍不限於此，資料是類別型的資料(categorical data)、無母數的資料(non-parametric)、非多元常態的資料都可以使用多重插補法，未來可就各種不同型態的資料進行多重插補法，並比較多重插補法在何種分布或者是何種類型的資料下較能得到穩健的分析結果。

在後續研究中，也可以針對同一種統計方法，如針對迴歸分析，使用不同資料進行多重插補法的分析，以便了解多重插補法在同一統計分析方法下，是否可得到一致的插補效果。

參考資料

一、中文參考資料

1. 李興南(2003)。在樣本完全闕失之多重插補方法的比較分析。台灣大學碩士論文。
2. 沈明來(2007)。實用多變數分析，第二版。九州圖書文物有限公司。
3. 沈明來(2007)。試驗設計學，第三版。九州圖書文物有限公司。
4. 吳東霖、林傑斌、劉明德(2003)。SAS 與統計模式建構。文魁資訊股份有限公司。
5. 黃齡葦(2005)。遺失資料之多重插補法模擬比較研究。台灣大學碩士論文。
6. 彭昭英(2009)。SAS 統計軟體研討會。國立台灣大學統計教學中心。
7. 楊棋全(2004)。指數與韋伯分佈遺失值之處理。國立中央大學碩士論文。

二、英文參考資料

1. Agresti, A. (1990), Categorical Data Analysis. New York: John Wiley & Sons, Inc.
2. Donald F. Morrison (1990), Multivariate Statistical Methods.3rd. McGraw-Hill International editions.
3. David S. Moore & George P. McCabe, Introduction to the practice of statistics, Forth Edition, W. H. Freeman and Company. New York.
4. Hotelling, H. (1936), Relations Between Two Sets of Variables, *Biometrika*, 28, 321 – 377
5. Johnson and Wichern(2007),Applied Multivariate Statistical Analysis,6th,Pearson
6. Patrick Royston. (2004)，Multiple imputation of missing values. *The Stata Journal* 4, Number 3, pp. 227–241
7. Rubin D.B.(1976), Inference and missing data, *Biometrika* 63:581-592.
8. Rubin D.B.(1987),Multiple Imputation for Nonresponse in Survey, Wiley
9. Sas Institute Inc.,(2002), SAS/STAT User's Guide, SAS Institute Inc., Cary, NC.
10. Schafer,J.L. (1997), Analysis of Incomplete Multivariate Data. New York: Chapman and Hall/CRC.
11. Schafer,J.L. (1999)，Multiple imputation : a primer, *Statistical Method in Medical Research* 8: 3-15
12. Spearman, C. (1904), "General Intelligence Objectively Determined and Measured," *American Journal of Psychology*, 15, 201 - 293.