國立台灣大學公共衛生學院流行病學與預防醫學研究所

碩士論文

Graduate Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University

Master Thesis

針對遺失變數的統計方法

Statistical Method for Missing Covariates

劉庭吟

Ting-yin Liu

指導教授：張淑惠 博士

Advisor: Shu-Hui Chang, Ph.D.

中華民國 101 年 7 月

July, 2012

國立臺灣大學碩士學位論文
口試委員會審定書

針對遺失變數的統計方法

Statistical method for missing covariates

本論文係 劉庭吟 君（學號 r99849011）在國立臺灣大學流行病學與預防醫學研究所完成之碩士學位論文，於民國 101 年 07 月 20 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：張淑惠 （簽名）

（指導教授）

戴政

鄭明華

陳秀熙

致謝

終於來到了這一頁，終於離畢業只剩最後一道線了，那種如釋重負的感覺，真是令人身心舒暢。在寫著這一頁的致謝，整個碩士生涯的一點一滴的在我的腦海中一一瀏覽過，記憶裡那些跑程式、讀論文的痛苦也似乎淡了些。論文能在後期的趕製中如期完成，這些都要歸功於給我信心與支持的強力後盾，撐著我走到最後，在此致上最真誠的感謝。

首先我要感謝指導教授張淑惠老師兩年來的栽培，謝謝她帶著我踏入更高的學識殿堂，還給了我充分的機會與空間，鼓勵、激勵我勇敢的往前行，讓我在學術上與生活上都留下許多深刻的成長與體驗。研究室的學長、學姐與同學是幫助我成長的重要人物。因此我要感謝最帥氣的宗仁學長，跟在他身邊讓我學會了很多研究上和生活上的智慧；感謝經常幫我打氣的同班同學們：郁雅、子庭、貴鈴、秋涵、若青與薇茹，陪伴我度過每一個五味雜陳的日子，一同享受碩士生的甘甜苦樂，彼此互相打氣與支持，互相體諒與幫忙；也要感謝同一組中最聰明的俊秀哥，他在生活上與人際溝通中，教會了我許多待人處事的道理，並給予我在學習上許多幫助。

最後，我要感謝我的家人為我提供一個溫暖的窩與避風港，讓我可以沈澱自己慌亂的情緒，謝謝爸爸、媽媽對我的寵愛和信任，讓我可以全心專注在自己的論文寫作上，不用為許多繁瑣的事情分心；謝謝姐姐們，在我心情不好時，總是擔任我的心情抒發桶，並提供我生活上的許多支助；感謝總是扮演搞笑角色的小弟，總是能讓我在苦悶時開懷大笑。

今天能走到最後，繳出一份令人滿意的成果，都要感謝這一路上曾幫助過我的師長、家人與朋友們，感謝你們，我將這份喜悅與你們分享，謝謝！

庭吟　於民國 101 年 7 月 30 日

摘要

在環境檢測與實驗研究中，由於收集的資料容易受儀器的偵測極限影響而產生遺失值。過去多數文獻只針對模型中至多兩個變數受偵測極限影響的資料，進行簡單替代法、插補法或模式建構法等的處理方法；而當模型中含有多個受偵測極限影響的變數且大多全為連續變數或類別變數時，則使用蒙地卡羅 EM 演算法搭配抽樣法以便解決高維度的積分問題，進而求得參數的估計。

本論文著重於羅吉斯模型分析中同時具有多個受偵測極限影響的連續變數和具有隨機遺失機制的類別變數而更為複雜的資料，並提供一個估計迴歸係數的估計方法。我們利用蒙地卡羅積分方法解決 EM 演算法中 E 步驟因受偵測極限與隨機遺失影響所產生的高維度積分。最後並引入已有的另兩種解決受偵測極限影響的資料之方法，比較不同方法的各自表現。

由模擬結果中，顯示本論文的方法在不同的設限比例下，迴歸係數的估計都較 Schistermane*t al.*(2006)與完整觀察個體分析來得不偏與精準。

關鍵字：偵測極限、最大概似估計量、隨機遺失、Monte Carlo EM、牛頓法

Abstract

In many environmental and laboratory studies, instrument detection limits often lead to missing values of the data. The existing methods for the regression analysis for the data with at most two covariates subject to detection limits include simple substitution, imputation, and model-based methods. While either multiple continuous covariates or multiple categorical covariates alone are subject to detection limits, the most common approaches are the model-based method, Expectation-Maximization (EM) algorithm, and a Monte Carlo version of EM algorithm to obtain the maximum likelihood estimates via sampling.

In this paper, we consider a more complex case of missing covariates that both multiple continuous covariates subject to detection limits and categorical covariates with missing at random mechanism are presented in the logistic regression analysis. The aim of this paper is to provide a method for estimating the parameters of regression models for data with covariates subject to detection limit and missing at random mechanism. We use the Monte Carlo version for the E-step of the EM algorithm to tackle the high dimensional integration and summation due to the missing covariates subject to detection limits and random missing. We conduct a simulation study to compare the performance of the proposed Monte Carlo EM algorithm approach with the complete-case method and the imputation method proposed by Schisterman et al. (2006).

The results of the simulation study showed that the proposed approach resulted in

relatively unbiased estimates with smaller standard error than the complete-case method and the imputation method by Schisterman et al, (2006).

Keywords : detection limits; maximum likelihood estimation; missing at random; Monte Carlo EM; Newton-Raphson methods

目錄

圖表目錄

第一章　緒論

第一節　研究背景

流行病學研究中，常收集環境檢測與實驗研究的資料，藉由相關的統計分析方法來測定某特定因子與研究事物的相關性，此階段分析的準確度必須仰賴其所收集的資料之完整性，以便進行較為嚴謹的推論。而在資料收集階段中，部分因子須在實驗室內透過各類精密儀器的檢測來得到其所能檢測的最小檢出結果，由於每個儀器有其檢測的偵測範圍，因此對此部分因子所收集的資料即會受儀器所能偵測的範圍影響。一般定義儀器的偵測範圍為偵測極限(limit of detection；LOD)，若偵測物質落於儀器的偵測極限外，則此物質的檢測結果通常標示為「未檢出」。例如：假設實驗室採用偵測極限為 2.5ppm 的三聚氰胺濃度檢測方法，若待測物質其檢測結果低於 2.5ppm，則檢驗人員無法武斷的斷定此物質「不含三聚氰胺」，亦不能確定其「含有三聚氰胺的濃度」，因此在檢測資料上只能標示「未檢出」來做為資料的呈現。若欲觀察的部份樣本資料因受到儀器的偵測極限所控制而被標示為「未檢出」，這種資料的呈現上可視為設限資料(censored data)，亦為一種不完整資料型態。當自變數或反應變數受偵測極限影響而有不完整資料時，直接此對資料做分析，則所估計出來的參數是否有偏差？其分析結論是否準確？為了解決由偵測極限所衍生出來的問題，過去研究學者提出許多不同的處理方法以進行資料分析。

完整觀察個體分析(complete-case analysis)是統計上最簡單且也最直覺的處理不完整資料的方法，它最早是用來處理數據中具有遺失資料的情況。這個方法是當觀察個體所收集的變項上若有觀察值落於偵測極限外時，則將觀察個體整筆資料移除。雖然在完全隨機遺失機制(missing completely at random mechanism；MCAR)的假設下，完整觀察個體分析可得到不偏的參數估計量，但是相對的，當設限資料過多時，統計分析上會損失過多的樣本資訊造成資訊的流失且亦會使估計量的

變異數隨著刪除的樣本數增加而變大。為了不損失過多樣本資訊，在統計上也有人利用偵測極限(LOD)的函數值來取代設限資料，例如：LOD 或 LOD/2，雖然這種方法既簡單又好操作，但利用此方法所估計出來的參數容易造成高度偏差。而 Richardson & Ciampi(2003)提出在簡單線性迴歸模型中，當只考慮單一變數受偵測極限影響且已知變數的母體分佈下，利用變數的條件期望值取代設限資料，即若假設一簡單迴歸模型$y = \alpha + \beta x + \varepsilon$，y 為疾病狀態變數(y＝1 代表有病，y＝0 代表沒病)，x 為一非負且連續暴露因子，ε為一隨機誤差且服從常態分佈，其平均值為 0，變異數為常數，則在假設偵測下限為 dl 的情形下，利用替代值$a = E(x_i | x_i < dl)$對低於檢測界限 dl 的設限資料進行取代，其中x_i為已知變數之母體分佈的理論值。Schisterman *et al.* (2006)則是提出利用已觀察到的樣本x_i^*之期望值取代設限資料，即利用替代值為$a^* = \frac{\sum_i x_i^* I(x_i^* \geq dl)}{\sum_i I(x_i^* \geq dl)}$進行取代。此兩種方法都可以得到不偏的參數估計量。在遺失資料的處理中，亦被廣泛應用的方法則是模式建構法，即利用最大概似(maximum likelihood method)理論來處理具有遺失值的資料；Nie *et al.* (2010)即利用模式建構法處理單一受偵測極限影響的變數，在已知變數分佈下比較最大概似法和簡易的替代法的參數估計量，其中以最大概似函數法所估計的參數具有較不偏的性質且其所估計出來的標準誤較簡易的替代法來得小。

若資料中有兩個受偵測極限影響的變數時，D'Anael *et al.*(2008)在 Cox 迴歸模式中提出 EM 演算法(Expectation - Maximization algorithm)的來解決此種問題。上述的解決方法皆是在最多兩個受偵測極限影響的變數下做討論，May, Ibrahim, & Chu (2011)則是探討在多個變數受偵測極限的影響下，提出可利用蒙地卡羅觀點的 EM 演算法來對含有設限資料的連續變數進行參數估計的迭代。

第二節　研究動機

在環境研究與實驗室的研究中，受儀器偵測極限而造成的設限資料是其不可或缺的資料型態之一，尤其當設限資料的比例過多時，若用較簡易的完整觀察個體分析，則會損失過多的樣本資訊，此時所估計出來的參數是否可信賴？因此，探討如何處理設限資料的不完整數據以減少其所帶來的資訊損失並增加估計的精準度有其重要性。插補法(imputation)中的 EM 法是解決這類問題最熟知的方法之一，但即使引入此方法，在估計上仍將面臨幾個困難點，其中包括：(1)當 EM 法中的「E(expectation)步驟」若沒有封閉型式，或有封閉型式但卻極為複雜時，則會增加「M(maximization)步驟」計算的困難度；(2)一般收集的資料中，有時會同時會收集到連續資料與類別型資料，當此兩類型的資料皆出現不完整數據且變數與變數間彼此都有相關時，連續與類別型資料各別進行 EM 法，則會忽略其變數間的相關性而造成其參數估計上的偏誤。因此處理 EM 法中 E 步驟的積分計算與處理連續變數與類別變數間的相關性都是需要進一步討論的問題。

過去處理不完整資料的方法中，Little & Schluchter(1985)針對遺失值資料提出利用最大概似估計法透過 EM 演算法來解決變數中同時含有連續變數與類別變數的問題。在本文中，將延續 May, Ibrahim, & Chu (2011)所提的方法，將偵測極限的資料中只有連續變數的情形延伸到同時含有受偵測極限影響的連續變數與具有隨機遺失機制的類別變數之情況下做探討。

第三節　研究目的

上節將主要的研究動機清楚說明後，本論文主要研究目的為：

(1)設立一組包含受偵測極限影響的連續變數之設限資料與隨機遺失的類別型資料之樣本，利用蒙地卡羅 EM 法得到最大概似估計量。

(2)計算在不同的設限比例下，本論文的方法和 Schisterman *et al.* (2006)與完整觀察個體分析在參數估計上的不偏性與有效性。最終希望能在不同的設限比例下，建議研究者在此類型的資料型態下應採用何者方法進行參數的估計較為適合。

第二章　文獻探討

第一節　遺失值機制

學者 Rubin(1976)一文中，已先介紹隨機遺失(missing at random)與隨機觀察(observed at random)的機制，而後 Little & Rubin (1987)定義出遺失值機制的三種類型：完全隨機遺失(missing completely at random；MCAR)、隨機遺失(missing at random；MAR)與非隨機遺失(not missing at random；NMAR)。假設一完整資料 $Y = (y_{ij})_{n\times m}$ 為一矩陣資料，且可被區分成Y_{obs}與Y_{mis}兩部份，其中Y_{obs}代表完整資料Y中可被觀察到的資料，Y_{mis}代表Y中所遺失的資料。定義遺失數據指標矩陣(missing data indicator matrix)$M = (m_{ij})_{n\times m}$，其中當y_{ij}為遺失值時，m_{ij} =1；當y_{ij}可被觀察到時，$m_{ij} = 0$。根據上述定義，我們可定意義：

(i)完全隨機遺失：當遺失值 M 既不依賴觀察值y_{obs}也不依賴遺失值y_{mis}，具有此性質的缺失機制則稱為完全隨機遺失。

(ii)隨機遺失：當遺失值僅和觀察值y_{obs}有關而與缺失值y_{mis}無關，表示遺失值之間彼此是獨立的，且資料的遺失受觀察值的影響，具有此性質的缺失機制稱為隨機遺失。

(iii)非隨機遺失：若遺失值彼此之間是有相關的，即遺失值依賴於y_{mis}時，則稱此缺失機制為非隨機遺失。換句話說，當數據資料的遺失機制非完全隨機遺失且也非隨機遺失時，此資料的遺失機制即稱為非隨機遺失。

就學者 Rubin 的定義，若遺失資料指標矩陣與遺失資料並無任何關係，在這種情形下，完全隨機遺失與隨機遺失機制又稱為可忽略的遺失資料機制(ignoring missing data mechanism)；而非隨機遺失因其資料所產生的遺失並非在研究者的控制下，因此不可忽略，故非隨機遺失機制又可稱為不可忽略遺失資料機制(nonignorable missing data mechanism)。

第二節　遺失值之處理方法

自 1950 年代開始，陸續有學者發表有關於處理遺失資料的文獻，如 Hartly(1958)一文中探討過去文獻為離散分佈的截切(truncated)資料與設限(censored)資料下，透過對資料的分佈做假設，提出利用最大概似估計法處理遺失值資料。Dempster, Laird, & Rubin (1977)提出了著名的 EM 演算法，在遺失機制為隨機遺失時，透過對資料做合理的分佈假設，利用 E 步驟的期望化階段來補足遺失的訊息，再進行概似函數最大化的 M 步驟來估計參數，藉由遞迴過程使估計量逐漸達到正確的數值。Rubin(1978)提出多重插補法(multiple imputation)的概念，是一種對資料進行數據擴充和統計分析的方法。它使用一系列共 M 個可能的數值來插補每個遺失值，產生 M 個完整數據資料，再對此 M 個完整數據資料進行一般正規的統計分析，然後結合這 M 個分析結果進行分析推斷。而後，Little & Rubin (1987)將遺失值的處理方法大致分成四個類別：

(1)完整的觀察個體分析(procedures based on completely recorded units)

(2)加權法(weighting procedures)

(3)插補法(imputation-based procedures)

(4)模式建構法(model-based procedures)

近年來大部分的統計研究廣泛的運用模式建構法，此方法主要利用概似最大化(maximum likelihood)的統計理論，事先預設某一母體的分佈模型，再以觀測所得之樣本透過概似最大化以求得參數估計量；即透過 ML 理論或者是 EM 理論，進行遺失值的處理，如 Dempster, Laird, & Rubin (1977)、Little & Schluchter(1985)等。本研究即是透過模式建構法，預先對母體作分佈假設，再透過蒙地卡羅 EM 演算法(Monte Carlo EM method)求得參數估計量。

當資料中出現連續資料與離散資料且部分資料為遺失狀態時，在統計方法處理上較一般的單一連續或離散資料還來得較複雜，下面將介紹 Little & Schluchter (1985)提出處理當資料同時含有類別和連續變項的情形下其遺失值的處理與參數

估計，以下將對此篇文章的資料結構做介紹。首先我們先對資料做變數的定義：

N：樣本數

X_s：含有 p 個連續變數的向量，即$X_s = [x_{1s}, x_{2s}, \cdots, x_{ps}]^T \quad s = 1,2 \cdots, N$

Y_s：含有 q 個類別變數的向量，即$Y_s = [y_{1s}, y_{2s}, \cdots, y_{qs}]^T s = 1,2 \cdots, N$

I_i：第 i 個類別變項內含有I_i個項目

C：q 維的列聯表細格數，即$C = \prod I_i$

$X_{s,obs}$：個體 s 所能觀察到的連續變數$s = 1,2 \cdots, N$

$X_{s,mis}$：個體 s 不能觀察到的連續變數$s = 1,2 \cdots, N$

E_m：只在第 m 個為 1，其餘為 0 的一$C \times 1$向量$m = 1,2 \cdots, C$

W_s：為一$C \times 1$的向量，當個體 s 在 q 維列聯表中只屬於某一細格時，

則$W_s = E_m$

完整資料(X_s, W_s)的分佈可藉由$(X_s|W_s)$的條件分佈與W_s的邊際分佈來表達。先從W_s的邊際分佈先看，由於 N 個個體會落於 q 維列聯表中的某一細格內，因此 N 個個體可看成服從多項式分佈，且每一細格機率為$\pi_m = pr(W_s = E_m)s = 1, \cdots, Nm = 1,2 \cdots, C$。而$W_s = E_m$代表所有的類別變項皆觀察到的情形下，故在給定$W_s = E_m$的條件下時，$X_s$則服從多變量常態分佈，其期望值為$\mu_m$，共變異數矩陣為$\sum$，即$X_s \sim N_p(\mu_m, \sum)$。若就列聯表中的細格來看，假設$\Gamma = [\mu_1, \mu_2, \cdots, \mu_C]$為一$p \times C$的矩陣，則$X_s$在給定$W_s$的條件分佈下又可寫成$N_p(\Gamma W_s, \sum)$。

因為$f(X_s|W_s, \Gamma, \sum) \sim N_p(\Gamma W_s, \sum)$，因此其機率函數可寫成

$$f(X_s|W_s, \Gamma, \sum) = f\left(x_{1s}, x_{2s}, \cdots, x_{ps} \middle| W_s, \Gamma, \sum\right)$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}} |\sum|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(X_s - \mu_m)^T \sum^{-1} (X_s - \mu_m)\right\}$$

$$= (2\pi)^{-\frac{p}{2}} |\sum|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X_s - \mu_m)^T \sum^{-1} (X_s - \mu_m)\right\} \quad (2.1)$$

則完整資料的 log 概似函數為：

$$\ell(\Gamma, \sum, \pi) = \log\{\prod_{s=1}^{N} f(X_s|W_s, \Gamma, \sum) f(W_s|\pi)\}$$

$$= \sum_{s=1}^{N}\{\log\ f(X_s|W_s,\Gamma,\sum) + \log f(W_s|\pi)\}$$

$$= -\frac{1}{2}N\{p\log(2\pi) + \log|\sum|\} - \frac{1}{2}tr(\sum^{-1}\sum_{s=1}^{N}X_s\,X_s^T) + tr\{\sum^{-1}\Gamma(\sum_{s=1}^{N}W_s\,X_s^T)\}$$

$$+\sum_{m=1}^{C}\left\{(\sum_{s=1}^{N}W_{sm})\left(\log\pi_m - \frac{1}{2}\mu_m^T\sum^{-1}\mu_m\right)\right\} \quad (2.2)$$

從(2.2)式可知完整資料的結構模式屬於正規的指數家族分佈(exponential family)，且其充分統計量(sufficient statistics)為$\sum_{s=1}^{N}X_s\,X_s^T$、$\sum_{s=1}^{N}W_s\,X_s^T$及$\sum_{s=1}^{N}W_{sm}$。EM 演算法的 E 步驟即在給定可觀察到的數值和目前所估計的參數值下對(2.2)式取條件期望值，則在第$(t+1)$次的迭代中，其 E 步驟以數學符號表達為：$E\left(\ell(\Gamma,\sum,\pi)\middle|X_{s,obs},S_s,\pi^{(t)},\Gamma^{(t)},\sum^{(t)}\right)$。由於完整資料屬於正規的指數家族分佈，則第$(t+1)$次的迭代的 E 步驟可改成對充分統計量取條件期望值，即

$$T_{1s} = E\left(X_sX_s^T\middle|X_{s,obs},S_s,\pi^{(t)},\Gamma^{(t)},\sum^{(t)}\right) \quad (2.3\text{-a})$$

$$T_{2s} = E\left(W_sX_s^T\middle|X_{s,obs},S_s,\pi^{(t)},\Gamma^{(t)},\sum^{(t)}\right) \quad (2.3\text{-b})$$

$$T_{3s} = E\left(W_s\middle|X_{s,obs},S_s,\pi^{(t)},\Gamma^{(t)},\sum^{(t)}\right) \quad (2.3\text{-c})$$

對(2.2)式分別對參數$\Gamma,\sum,\pi$做微分可得參數$\Gamma,\sum,\pi$各自的最大概似估計量：

$$\hat{\pi} = N^{-1}\sum W_s \quad (2.4\text{-a})$$

$$\hat{\Gamma} = (\sum X_s\,W_s^T)(\sum W_s\,W_s^T)^{-1} \quad (2.4\text{-b})$$

$$\hat{\sum} = N^{-1}\sum\left(X_s - \hat{\Gamma}W_s\right)\left(X_s - \hat{\Gamma}W_s\right)^T \quad (2.4\text{-c})$$

則 EM 演算法的 M 步驟即是對(2.4-a)式到(2.4-c)式中的充分統計量$\sum X_sX_s^T$、$\sum W_sX_s^T$與$\sum W_s$分別以(2.3-a)式到(2.3-c)式的條件期望值T_{1s}、T_{2s}與T_{3s}取代，因此第$(t+1)$次迭代中 EM 演算法其最大概似估計量為

$$\hat{\pi} = N^{-1}\sum T_{3s}$$

$$\hat{\Gamma} = (\sum T_{2s})^T(D)^{-1}$$

$$\hat{\sum} = N^{-1}\{\sum T_{1s} - (\sum T_{2s})^TD^{-1}(\sum T_{2s})\}$$

其中 D 為一對角線矩陣，其除了主對角線為$\sum T_{3s}$內的元素外，其餘皆為 0。

第三節　受偵測極限影響的資料及其處理

在實驗室或者是環境檢測中，化學分析是一門重要的學門，為了能了解某物質的特性，通常需藉由儀器的使用以便達到分析化學的目的。根據 MacDougall & Daniel(1980)的定義，我們又可透過分析化學的分析目的主要分成兩類：定性分析(qualitative analysis)與定量分析(quantitative analysis)。定性分析是指測定某待測物的物質性質，即分析某待測物中所含的成分種類以便確定此物質的特性或者結構鑑定(identification)；定量分析則是進一步確定成分的含量，分析某一欲測成分在試樣中所佔的比例或濃度。而在此兩種分析方法上，其儀器的使用上皆有其偵測極限(detection limit)的範圍，分別為方法偵測極限(method detection limit)與儀器偵測極限(instrument detection limit)。方法偵測極限是指一待測物在某一種基質中所能檢測得之最小濃度，又稱為定性極限；而儀器偵測極限是指待測物之最小或最低量的濃度(訊號)，使儀器能夠在偵測時，產生一可與空白訊號區別的訊號者，又可稱為定量極限。我們一般在實驗室中或環境檢測記錄中，通常所欲測定的是待測物的最小濃度，因此收集的資料容易受儀器的偵測極限所影響而產生一左設限(left-censored)或右設限(right -censored)的資料。

處理上述因偵測極限影響而產生的右設限或左設限資料結構，Lynn (2001)一文中探討血漿中 HIV RNA 測量的左設限資料使否與其病毒載量的量化有關，由於資料的遺失會與未能觀察到的數值有相關，由此可知此資料的遺失機制為非隨機遺失，故在統計方法上較不適用多重插補法或其他的替代(substitution)方法，而是利用模式建構法，以概似最大化(maximum likelihood)來估計參數，並比較在不同設限資料比例下，替代法、多重插補法與其文中所使用的模式建構法在參數估計上精準度，其中以模式建構法所估計出的參數較為不偏與精準。當資料中的某一待檢測變數(暴露因子)有測量誤差時，Richardson & Ciampi (2003)一文中提出在已假設資料的分佈且其暴露因子服從對數常態分佈或 Gamma 分佈的情況下，在條件於已觀察到的資料，利用此變數的條件期望值做為遺失資料的替代數值，並將暴

露因子有測量誤差與無測量誤差的參數估計值之偏差做比較，發現在有假定有測量誤差時，其參數估計值的偏差較小。Lubin *et al.* (2004)則是利用分佈抽樣的概念，在對資料做適合的分佈假設時，比較簡單的替代法(如偵測極限的二分之一倍)和對適合的分佈做隨機抽樣並填入(fill-in)遺失資料的方法；由結論可知在設限資料比例不超過 30%下，相較於簡單的替代法，填入法可得到不偏的參數估計值，但當設限資料比例超過 30%後，填入法的參數變異數估計將會產生較大的偏差。Schisterman *et al.*(2006)一文改進 Richardson & Ciampi (2003)文章中須對資料分佈的設立，提出無須對共變數做分佈的假設之處理方法，若此共變數有一低的檢測門檻(lower threshold)，則利用可觀察到的資料之平均值做為遺失資料的替代值，即計算所有大於檢測門檻的數值之平均數做為替代值；此方法所得到的參數估計仍為不偏的，且其參數變異數估計較替代法來的小。而處理 Cox 迴歸模型(Cox regression model)具有設限資料的變數，D'Anaelo *et al.* (2008)提出改進 Rigobon & Stoker (2004)的指標方法(index method)建立 Cox 迴歸模型並最多處理兩個設限變數(censored covariates)的情形，透過比較完整觀察個體分析、替代方法與指標方法，可知指標方法在參數估計上的不偏性與精準度較其他方法來得好。

上述的方法都在只有單一受偵測極限影響的變數時或最多兩個設限變數的情形下做討論。在多變數資料為遺失的情形下，Stubbendick & Ibrahim (2006)提出利用模式建構法處理在長期追蹤的離散資料下，非隨機遺失之變數的參數估計，並透過蒙地卡羅 EM(Monte Carlo EM)演算法解決高維度積分。而 May 等人(2011)則是提出處理多個受偵測極限影響的變數，並用蒙地卡羅 EM 演算法透過 ARMS(adaptive rejection metropolis sampling)抽樣法以解決高維度積分，並獲得最大概似估計量。下面將介紹 May, Ibrahim, & Chu (2011)於論文中所提出的方法：

設有$\{(x_i, y_i); i = 1, \cdots, n\}$，共 n 個獨立的樣本觀察值，y代表反應變數，x代表含有 p 個變數(covariates)的向量。在這裡，我們把$p(y_i|x_i, \theta)$定義為條件分佈$[y_i|x_i]$的條件機率，θ為一$k \times 1$的參數向量；在 GLM 中，通常$\theta = (\beta, \gamma)$，β代表迴歸模式

中的迴歸係數，γ而代表分散參數(dispersion parameter)；$p(x_i|\alpha)$為$[x_i]$的邊際機率，其中α為邊際分佈函數$[x_i]$所對應的參數。因此，(y_i, x_i)的聯合分佈函數可寫成一序列的一維條件分佈函數：

$$p(x_i, y_i) = p(y_i|x_i, \theta)p(x_i|\alpha)$$

結合上述的(x_i, y_i)的聯合分佈公式，則完整數據的 log 概似函數可寫成如下：

$$\log - \text{likelihood } l(x, y|\gamma) = \log \prod_{i=1}^{n} p(x_i, y_i)$$

$$= \sum_{i=1}^{n} \log[p(y_i|x_i, \theta)p(x_i|\alpha)] = \sum_{i=1}^{n} \log[p(y_i|x_i, \theta)] + \log[p(x_i|\alpha)] \quad (2.5)$$

其中$\gamma = (\theta, \alpha)$，$\theta$為我們主要有興趣的參數。

為了能探討設限變數的情況，令x_i為包含$x_{cens,i}$與$x_{obs,i}$兩部分，$x_{obs,i}$代表完整觀察到的變數(fully observed covariates)，$x_{cens,i}$代表一$q_i \times 1$的設限變數向量(vector of censored covariates)。對第 i 個 subject 的第 j 個變數(covariate)而言，其設限區間(censoring interval)為(c_{lij}, c_{uij})，且只考慮所有設限區間皆已知的情形。定義$(c_l < x_{cens,i} < c_u)$為$x_{cens,i}$中的變數所對應的設限區間，即

$$(c_l < x_{cens,i} < c_u) \equiv \bigcap_{x_{ij} \in x_{cens,i}} (c_{lij} < x_{ij} < c_{uij}) \quad (2.6)$$

藉由上述的定義，即可來處理有關受偵測極限影響的變數。為了能方便闡述，這邊的分散參數指考慮τ為 1 的情形，即$\theta = \beta$。

$$Q(\gamma|\gamma^{(t)}) = E(\text{完整數據的 log 概似函數}|\text{可觀察到的觀察個體資料})$$

$$= E(l(x_i, y_i|\gamma)|\text{第 i 個觀察個體之觀察值})$$

$$= E\{\sum_{i=1}^{n} \log[p(y_i|x_i, \beta)] + \log[p(x_i|\alpha)] \,|\, \text{第 i 個觀察個體之觀察值}, \gamma^{(t)}\}$$

$$= \sum_{i=1}^{n} E\{\log[p(y_i|x_i, \beta)] + \log[p(x_i|\alpha)]|\text{第 i 個觀察個體之觀察值}, \gamma^{(t)}\}$$

值得注意的是，這邊的觀察個體之觀察值指的是能直接被觀察到的$(y_i, x_{obs,i})$資料和$(c_l < x_{cens,i} < c_u)$的資訊。因此，若假設設限變數是連續行變數時，則在 EM 演算法的 E 步驟在計算上為積分的型態，故可以把第 i 個觀察個體第 t 次迭代的 E 步驟寫成

$$Q_i(\gamma|\gamma^{(t)}) = \int \log[p(y_i|x_i, \beta)]\, p(x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}, c_l < x_{cens,i} < c_u) dx_{cens,i}$$

$$+\int \log[p(x_i|\alpha)]\, p\left(x_{cens,i}\middle|x_{obs,i}, y_i, \gamma^{(t)}, c_l < x_{cens,i} < c_u\right)dx_{cens,i}$$

$$=\int \log[p(y_i|x_i,\beta)]\, p\left(x_{cens,i}\middle|x_{obs,i}, y_i, \gamma^{(t)}\right) \times I\left(c_l < x_{cens,i} < c_u\right)dx_{cens,i}$$

$$+\int \log[p(x_i|\alpha)]\, p\left(x_{cens,i}\middle|x_{obs,i}, y_i, \gamma^{(t)}\right) \times I\left(c_l < x_{cens,i} < c_u\right)dx_{cens,i}$$

$$= Q_i^{(1)}\left(\beta|\gamma^{(t)}\right) + Q_i^{(2)}\left(\alpha|\gamma^{(t)}\right) \quad (2.7)$$

由上式的$Q_i\left(\gamma|\gamma^{(t)}\right)$可看出即使當$x_{cens,i}$所含的設限變數很少，在積分上其封閉型式(closed-form)的解仍會非常複雜，因此在此步驟的積分計算則採用蒙地卡羅的 EM 演算法來解決此方面的問題。為了計算在蒙地卡羅 EM 演算法 E 步驟的第$(t+1)$次迭代之(2.7)式，必須先產生一組來自條件截切分佈$\left[x_{cens,i}\middle|x_{obs,i}, y_i, \gamma^{(t)}\right]I\left(c_l < x_{cens,i} < c_u\right)$的樣本，此抽樣方法可以透過 ARMS 抽樣法對此條件截切分佈抽樣。對第 i 個觀察個體而言，每一次的抽樣個數即為第 i 個觀察個體的設限變數個數q_i，換句話說，抽樣樣本z_i為一包含q_i個樣本點的向量。透過蒙地卡羅 EM 演算法，E 步驟的第$(t+1)$次迭代為：

$$Q_i\left(\gamma|\gamma^{(t)}\right) = \frac{1}{m_i}\sum_{k=1}^{m_i} \ell\left(z_{ik}, x_{obs,i}, y_i, \gamma\right) = Q_i^{(1)}\left(\beta|\gamma^{(t)}\right) + Q_i^{(2)}\left(\alpha|\gamma^{(t)}\right) \quad (2.8)$$

且$Q\left(\gamma|\gamma^{(t)}\right) = \sum_{i=1}^{n} Q_i\left(\gamma|\gamma^{(t)}\right)$，因此 EM 演算法的 M 步驟為對(2.8)式做一階微分並令為零求解，即

$$\dot{Q}(\gamma|\gamma^{(t)}) = \sum_{i=1}^{n} \dot{Q}_i\left(\gamma|\gamma^{(t)}\right) = \sum_{i=1}^{n} \frac{1}{m_i}\sum_{k=1}^{m_i} \frac{\partial}{\partial\gamma}\ell\left(z_{ik}, x_{obs,i}, y_i, \gamma\right) = 0 \quad (2.9)$$

透過不斷的迭代，當滿足$\left\|\beta^{(t+1)} - \beta^{(t)}\right\| < \varepsilon$的條件時，則可得到參數$\beta$的估計即為$\beta^{(t+1)}$。

第三章　研究方法

在實驗室測量試驗中，往往資料會受測量儀器的偵測極限影響，而使資料有遺失值的產生，當資料中的連續變數受偵測極限所影響且類別變數具有隨機遺失機制時，資料的分析更顯得複雜。在處理此類型的不完整資料時，我們考慮使用 EM 演算法在模型中找尋參數的最大概似估計量。因此，我們先建立一組完整的資料$\{(y_i, \mathbf{x}_i^T, \mathbf{z}_i); i = 1,2,\cdots,N\}$，在資料符號定義中，令$y_i$為第 i 個觀察個體的二元反應變項觀察值，而每一$y_i$所對應的$\mathbf{x}_i^T$為一$p \times 1$的連續共變數向量$(x_{i1}, x_{i2}, \cdots, x_{ip})$，$x_{ij}$代表第 i 個觀察個體的第 j 個連續變數觀察值，$j = 1,\cdots,p$；$\mathbf{z}_i^T$為$q \times 1$的類別變數向量$(z_{i1}, z_{i2}, \cdots, z_{iq})$，且令第 i 個體的第 j 個類別變數$z_{ij}$中含有$I_j$個項目(levels)。這q個類別變數會構成一q維且含有$C = \prod_j^q I_j$個細格數的列聯表，每一個個體會落於這C個細格數中。假設來自不同個體的共變數樣本彼此互相獨立，且來自同一個體的共變數樣本具有相關性，即$Cov(x_{ij}, x_{ik}) \neq 0$、$Cov(z_{ij}, z_{ik}) \neq 0$，且當$i \neq l$時，$Cov(x_{ij}, x_{lk}) = 0$、$Cov(z_{ij}, z_{lk}) = 0$。

在廣義線性模型(Generalized linear models)的假設下，我們可知道y_i服從一指數家族分佈(Exponential family)，並且可透過鏈結函數(link function)建立線性預測因子(linear predictor)與分佈期望值之關係。因反應變項y_i為二元資料，故考慮 logit 的鏈結函數可建立(3.1)式的模式：

$$g(E(Y_i)) = \text{logit}\, P(y_i|x_i, z_i) = \beta_{(p)}\mathbf{x}_i^T + \beta_{(q)}\mathbf{z}_i^T \qquad (3.1)$$

其中$\beta_{(p)} = (\beta_1,\cdots,\beta_p)$，$\beta_{(q)} = (\beta_{p+1},\cdots,\beta_{p+q})$皆為參數向量。

欲建立$(y_i, \mathbf{x}_i, \mathbf{z}_i)$的聯合分佈，則可透過$y_i$條件於$x_i$與$z_i$的條件分佈$[y_i|x_i,\mathbf{z}_i]$和$\mathbf{x}_i$條件於$\mathbf{z}_i$的條件分佈$[\mathbf{x}_i|\mathbf{z}_i]$與$\mathbf{z}_i$的邊際分佈$[\mathbf{z}_i]$而得。令$y_i$的條件分佈$[y_i|\mathbf{x}_i, \mathbf{z}_i]$其條件機率密度函數為$p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta)$，$\beta$為一$(p+q) \times 1$維的參數向量；$\mathbf{x}_i$的條件分佈$[\mathbf{x}_i|\mathbf{z}_i]$其條件機率密度函數為$p(\mathbf{x}_i|\mathbf{z}_i, \mu)$，$\mu$為$x_i$的條件分佈所對應的參數，因此可知道$\mathbf{x}_i$與$\mathbf{z}_i$的關係可透過參數$\mu$來決定；$\mathbf{z}_i$的邊際分佈$[\mathbf{z}_i]$其條件機率密度函數為

$p(\mathbf{z}_i|\pi)$，π為$\mathbf{z}_i$的邊際分佈所對應的參數。根據上述定義，$(y_i, \mathbf{x}_i, \mathbf{z}_i)$的聯合分佈可寫成：

$$p(y_i, \mathbf{x}_i, \mathbf{z}_i) = p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta)p(\mathbf{x}_i|\mathbf{z}_i, \alpha)p(\mathbf{z}_i|\pi) \quad (3.2)$$

且完整資料的對數概似函數(log likelihood)可寫成：

$$\ell(\beta, \alpha, \pi|y_i, \mathbf{x}_i, \mathbf{z}_i, i = 1, \cdots, N) = \log \prod_{i=1}^{N} p(y_i, \mathbf{x}_i, \mathbf{z}_i)$$

$$= \sum_{i=1}^{N}\{\log[p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta)] + \log[p(\mathbf{x}_i|\mathbf{z}_i, \alpha)] + \log[p(\mathbf{z}_i|\pi)]\} \quad (3.3)$$

因為主要有興趣的參數為迴歸係數β，因此其他參數α與π則為所謂的 nuisance 參數。

引用 Little & Schluchter(1985)對連續變數與類別變數之間的相關性描述，定義E_m為一$C \times 1$向量，其第 m 個值為 1，其餘為 0，$m = 1,2 \cdots, C$；Z_i為一$C \times 1$的向量，當第 i 個個體在 q 維列聯表中只屬於某一細格時，則$Z_i = E_m$。根據上述對連續資料與類別資料的符號定義下，其完整資料的概似函數可由下列 3 個邊際分佈與條件分佈的模式假設來表達：

(1) 由於 N 個觀察個體會落於 q 維列聯表的某一細格內，故 N 個觀察個體可視為服從一多項式分佈，且每一細格的機率為$\pi_m = pr(Z_i = E_m)$　$i = 1,2 \cdots, Nm = 1,2 \cdots, C$；即若設變數$n_m$為落於第 m 個細格內的觀察個數，$m = 1, \cdots, C$，且$\sum_{m=1}^{C} n_m = N$，則$(n_1, n_2, \cdots, n_C) \sim$Mutinomial$(N, \pi_1, \pi_2, \cdots, \pi_C)$

(2) 條件於第i個觀察個體落於q維列聯表中的第m個細格 (即$Z_i = E_m$)的情形下，共變數$\mathbf{x}_i$服從多變量常態$N_p(\mu_m, \Sigma)$，其中期望值μ_m為透過第 m 個細格數所決定；即$\mathbf{x}_i|Z_i = E_m \sim N_p(\mu_m, \Sigma)$。

(3) 在條件於$\mathbf{x}_i$與$\mathbf{z}_i$下，反應變數$y_i|\mathbf{x}_i, \mathbf{z}_i \sim$Ber$(p_i)$，其中成功機率$p_i$的定義為

$$p_i = p(y_i = 1|\mathbf{x}_i, \mathbf{z}_i) = \frac{\exp\{\beta_{(p)}\mathbf{x}_i^T + \beta_{(q)}\mathbf{z}_i^T\}}{1+\exp\{\beta_{(p)}\mathbf{x}_i^T + \beta_{(q)}\mathbf{z}_i^T\}}$$。

根據上述對資料的模式假設，其完整資料的概似函數為：

$$L(\beta, \alpha, \pi|y_i, \mathbf{x}_i, \mathbf{z}_i) = \prod_{i=1}^{N} p(y_i, \mathbf{x}_i, \mathbf{z}_i) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta)p(x_i|\mathbf{z}_i, \mu_m)p(\mathbf{z}_i|\pi_m)$$

$$= \prod_{i=1}^{N} \left\{ \left(\frac{\exp\{\beta_{(p)}\mathbf{x}_i^T + \beta_{(q)}\mathbf{z}_i^T\}}{1+\exp\{\beta_{(p)}\mathbf{x}_i^T + \beta_{(q)}\mathbf{z}_i^T\}} \right)^{I\{y_i=1\}} \left(\frac{1}{1+\exp\{\beta_{(p)}\mathbf{x}_i^T + \beta_{(q)}\mathbf{z}_i^T\}} \right)^{I\{y_i=0\}} \times \right.$$

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_m)^T\Sigma^{-1}(\mathbf{x}_i-\mu_m)\right\}\times\prod_{m=1}^{C}\pi_m{}^{z_{im}}\Bigg\}$$

$$\propto\prod_{i=1}^{N}\Bigg\{\left(\frac{\exp\{\beta_{(p)}\mathbf{x}_i^T+\beta_{(q)}\mathbf{z}_i^T\}}{1+\exp\{\beta_{(p)}\mathbf{x}_i^T+\beta_{(q)}\mathbf{z}_i^T\}}\right)^{I\{y_i=1\}}\left(\frac{1}{1+\exp\{\beta_{(p)}\mathbf{x}_i^T+\beta_{(q)}\mathbf{z}_i^T\}}\right)^{I\{y_i=0\}}$$

$$\times|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_m)^T\Sigma^{-1}(\mathbf{x}_i-\mu_m)\right\}\times\prod_{m=1}^{C}\pi_m{}^{z_{im}}\Bigg\} \qquad (3.4)$$

當資料的連續變項含有受偵測極限(detection limit)影響的變數時，我們令共變數$\mathbf{x}_i$包含兩$\mathbf{x}_{det,i}$與$\mathbf{x}_{obs,i}$兩部分，其中$\mathbf{x}_{det,i}$為一$r_i\times 1$的變數向量，代表第 i 個觀察個體含有r_i個受偵測極限影響而有遺失資料的變數，即$\mathbf{x}_{det,i}=\left(x_{i1}^*,x_{i2}^*,\cdots,x_{ir_i}^*\right)$；$\mathbf{x}_{obs,i}$為$(p-r_i)\times 1$向量，代表變數中的觀測值可被觀察到。假設對第 i 個 subject 的第 j 個變數(covariate)而言，其偵測極限的上限(upper bound)與下限(lower bound)分別為c_{lij}與c_{uij}，即設限區間(censoring interval)為(c_{lij},c_{uij})。定義$(c_l<x_{det,i}<c_u)$為$x_{det,i}$中的變數所對應的設限區間，即

$$(c_l<x_{det,i}<c_u)\equiv\bigcap_{x_{ij}\in x_{det,i}}(c_{lij}<x_{ij}<c_{uij})$$

而令含有遺失資料的類別變數$\mathbf{z}_i$亦包含$\mathbf{z}_{obs,i}$與$\mathbf{z}_{mis,i}$兩部分，其中$\mathbf{z}_{mis,i}$為第 i 個觀察個體有遺失資料的$s_i\times 1$之變數向量且具有隨機遺失機制(missing at random mechanism)，而$\mathbf{z}_{obs,i}$為代表可完全觀察到數值的$(q-s_i)\times 1$之變數向量。

當資料為完整數據資料(即沒有任何缺失值)時，我們可以直接對概似函數(3.4)式取對數，對其做一階微分並令為零以求有興趣參數的最大概似估計量(MLE)；但由於資料具有部分缺失資料，因此我們可以透過 EM 法(Dempster, Laird, & Rubin, (1977)的 E 步驟來加以補足遺失的資料，其 E 步驟為在給定所能觀察的資料下，對完整資料的對數概似函數取期望值，即(3.4)式的 EM 之 E 步驟可寫成：

$$Q(\gamma|\gamma^{(t)})=E\left\{\sum_{i=1}^{N}\log[p(y_i,\mathbf{x}_i,\mathbf{z}_i)]\,\middle|\,\text{observed}_i,\gamma^{(t)}\right\}$$

$$=E\left\{\sum_{i=1}^{N}\log[p(y_i|\mathbf{x}_i,\mathbf{z}_i,\beta)]+\log\left[p\left(\mathbf{x}_i\middle|\mathbf{z}_i,\mu_m\right)\right]+\log[p(\mathbf{z}_i|\pi_m)]\,\middle|\,\text{observed}_i,\gamma^{(t)}\right\}$$

$$=E\left\{\sum_{i=1}^{N}I\{y_i=1\}\log\left(\frac{\exp\{\eta_i\}}{1+\exp\{\eta_i\}}\right)+\sum_{i=1}^{N}I\{y_i=0\}\log\left(\frac{1}{1+\exp\{\eta_i\}}\right)\right.$$

$$\left.-\frac{1}{2}\log|\Sigma|+\frac{1}{2}(\mathbf{x}_i-\mu_m)^T\Sigma^{-1}(\mathbf{x}_i-\mu_m)+\sum_{m=1}^{C}\left(\sum_{i=1}^{N}Z_{im}\right)\pi_m\,\middle|\,\text{observed}_i,\gamma^{(t)}\right\} \quad (3.5)$$

其中，observed_i代表第 i 個個體中可被觀察到的變數之資料，$\gamma^{(t)}$代表在 EM 法中的第 t 次迭代之參數$\gamma^{(t)} = \left(\beta^{(t)}, \mu_m^{(t)}, \pi_m^{(t)}\right)$，$\eta_i = \beta_{(p)}\mathbf{x}_i^T + \beta_{(q)}\mathbf{z}_i^T$。不同於一般遺失資料的情況，對於遺失資料來說，我們並不能得到額外的遺失資訊，因此在第 i 個體的 E 步驟中所能觀察到的資訊僅僅為$(y_i, \mathbf{x}_{obs,i}, z_{obs,i})$，而在實驗室中所得的偵測極限資料，能知道其資料的遺失是由於其儀器的偵測極限範圍所導致，因此所能觀察到的資訊為$(y_i, x_{obs,i}, z_{obs,i})$與$(c_l < x_{det,i} < c_u)$。因此(3.5)式可寫成：

$$Q(\gamma|\gamma^{(t)}) = E\{\sum_{i=1}^{N} \log[p(y_i, \mathbf{x}_i, \mathbf{z}_i)] \,|\mathbf{x}_{obs,i}, y_i, \mathbf{z}_{obs,i}, \gamma^{(t)}, c_l < \mathbf{x}_{det,i} < c_u\}$$

故$Q_i(\gamma|\gamma^{(t)})$

$$= \sum_{\mathbf{z}_{mis,i}} \int \log[p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta)]\, p(\mathbf{x}_{det,i}, \mathbf{z}_{mis,i}|\mathbf{x}_{obs,i}, y_i, \mathbf{z}_{obs,i}, \gamma^{(t)}) \times I(c_l < \mathbf{x}_{det,i} < c_u) d\mathbf{x}_{det,i}$$

$$+ \sum_{\mathbf{z}_{mis,i}} \int \{\log[p(\mathbf{x}_i|\mathbf{z}_i, \alpha)] + \log[p(\mathbf{z}_i|\tau)]\}\, p(\mathbf{x}_{det,i}, \mathbf{z}_{mis,i}|\mathbf{x}_{obs,i}, y_i, \mathbf{z}_{obs,i}, \gamma^{(t)}) \times I(c_l < \mathbf{x}_{det,i} < c_u) d\mathbf{x}_{det,i}$$

$$= Q_i^{(1)}(\beta|\gamma^{(t)}) + Q_i^{(2)}(\mu_m, \pi_m|\gamma^{(t)}) \quad (3.6)$$

其中$p(\mathbf{x}_{det,i}, \mathbf{z}_{mis,i}|\mathbf{x}_{obs,i}, y_i, \mathbf{z}_{obs,i}, \gamma^{(t)})$為

$$p(\mathbf{x}_{det,i}, \mathbf{z}_{mis,i}|\mathbf{x}_{obs,i}, y_i, \mathbf{z}_{obs,i}, \gamma^{(t)})$$

$$= \frac{p(y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}, \gamma^{(t)})}{\sum_{\mathbf{z}_{mis,i}} \int p(y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}, \gamma^{(t)}) d\mathbf{x}_{det,i}} \quad (3.7)$$

在(3.6)式中，隨著受偵測極限影響的變數與隨機遺失的類別變數個數增多時，所要計算的期望值函數為多維的多重積分，並沒有有效的方法可以直接計算此式，而(3.7)式中的分母為$p(y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{obs,i}, \mathbf{z}_{mis,i}, \gamma^{(t)})$的邊際聯合密度函數，為針對遺失資料的變數$\mathbf{x}_{det,i}$與$\mathbf{z}_{mis,i}$做積分，亦為一難以計算的高維度積分，因此我們可以使用 Wei & Tanner(1990)所提的 MCEM 演算法利用蒙地卡羅積分(Monte Carlo Intergration)去近似(3.6)式中 EM 演算法的 E 步驟；即若假設欲求某函數h(x)的期望值$I = E_{f(x)}[h(x)] = \int h(x)f(x)dx$，則可從分佈f(x)中隨機產生M個樣本$(x_1, \cdots, x_M)$，即$(x_1, \cdots, x_M) \sim f(x)$，根據弱大數法則$I = \frac{1}{M}\sum_{k=1}^{M} h(x_k) \xrightarrow{p} E_{f(x)}[h(x)] = I$。為了計算

EM 演算法中的 E 步驟，故我們必須對每一個體i具有遺失資料的變數產生一組 M 個來自截切(truncated)分佈$\left[\mathrm{x}_{\mathrm{det,i}},\mathbf{z}_{\mathrm{mis,i}}\middle|\mathrm{x}_{\mathrm{obs,i}},\mathrm{y}_{\mathrm{i}},\mathbf{z}_{\mathrm{obs,i}},\gamma^{(\mathrm{t})}\right]\times\mathrm{I}\left(\mathrm{c}_{\mathrm{l}}<\mathbf{x}_{\mathrm{det,i}}<\mathrm{c}_{\mathrm{u}}\right)$的樣本點$\left(\mathbf{x}_{\mathrm{det,i}},\mathbf{z}_{\mathrm{mis,i}}\right)$，此抽樣可透過對下列的滿條件分佈(full conditional distribution)進行 ARMS 抽樣(adaptive rejection metropolis sampling)來完成：

$$\left[\mathbf{x}_{\mathrm{det,i}}\middle|\mathrm{y}_{\mathrm{i}},\mathbf{x}_{\mathrm{obs,i}},\mathbf{z}_{\mathrm{mis,i}},\mathbf{z}_{\mathrm{obs,i}},\gamma^{(\mathrm{t})}\right]\mathrm{I}\left(\mathrm{c}_{\mathrm{l}}<\mathbf{x}_{\mathrm{det,i}}<\mathrm{c}_{\mathrm{u}}\right)$$
$$\propto\left[\mathrm{y}_{\mathrm{i}}\middle|\mathbf{x}_{\mathrm{i}},\mathbf{z}_{\mathrm{i}},\gamma^{(\mathrm{t})}\right]\times\left[\mathbf{x}_{\mathrm{mis,i}}\middle|\mathbf{x}_{\mathrm{obs,i}},\mathbf{z}_{\mathrm{i}}\right]\mathrm{I}\left(\mathrm{c}_{\mathrm{l}}<\mathbf{x}_{\mathrm{det,i}}<ub\right)\quad(3.8)$$

$$\left[\mathbf{z}_{\mathrm{mis,i}}\middle|\mathrm{y}_{\mathrm{i}},\mathbf{x}_{\mathrm{obs,i}},\mathbf{x}_{\mathrm{det,i}},\mathbf{z}_{\mathrm{obs,i}},\gamma^{(\mathrm{t})}\right]\mathrm{I}\left(\mathrm{c}_{\mathrm{l}}<\mathbf{x}_{\mathrm{det,i}}<\mathrm{c}_{\mathrm{u}}\right)$$
$$\propto\left[\mathrm{y}_{\mathrm{i}}\middle|\mathbf{x}_{\mathrm{i}},\mathbf{z}_{\mathrm{i}},\gamma^{(\mathrm{t})}\right]\times\left[\mathbf{x}_{\mathrm{mis,i}}\middle|\mathbf{x}_{\mathrm{obs,i}},\mathbf{z}_{\mathrm{i}}\right]\times\left[\mathbf{z}_{\mathrm{mis,i}}\middle|\mathbf{z}_{\mathrm{obs,i}}\right]\mathrm{I}\left(\mathrm{c}_{\mathrm{l}}<\mathbf{x}_{\mathrm{det,i}}<ub\right)\quad(3.9)$$

藉由上述的抽樣方法，對每個觀察個體 i，我們可以得到一組樣本數為 M 的$\mathbf{v}_{\mathrm{i}}$樣本向量，$\mathbf{v}_{\mathrm{i}}=\left(\mathbf{x}_{\mathrm{det,i}},\mathbf{z}_{\mathrm{mis,i}}\right)$，因此第 i 個觀察個體其 EM 演算法的第$(\mathrm{t}+1)$次迭代之 E 步驟可寫成：

$$\mathrm{Q_i}\left(\gamma\middle|\gamma^{(\mathrm{t})}\right)=\frac{1}{\mathrm{M}}\sum_{\mathrm{k}=1}^{\mathrm{M}}\ell\left(\mathbf{v}_{\mathrm{ik}},\mathbf{x}_{\mathrm{obs,i}},\mathbf{z}_{\mathrm{obs,i}},\mathrm{y}_{\mathrm{i}}\right)$$
$$=\frac{1}{\mathrm{M}}\sum_{\mathrm{k}=1}^{\mathrm{M}}\mathrm{Q_i}^{(1)}\left(\beta\middle|\gamma^{(\mathrm{t})}\right)+\frac{1}{\mathrm{M}}\sum_{\mathrm{k}=1}^{\mathrm{M}}\mathrm{Q_i}^{(2)}\left(\mu_{\mathrm{m}},\pi_{\mathrm{m}}\middle|\gamma^{(\mathrm{t})}\right)$$

其中$\mathrm{Q}^{(1)}\left(\beta\middle|\gamma^{(\mathrm{t})}\right)$與$\mathrm{Q}^{(2)}\left(\mu_{\mathrm{m}},\pi_{\mathrm{m}}\middle|\gamma^{(\mathrm{t})}\right)$近似於：

$$\widehat{\mathrm{Q}}^{(1)}\left(\beta\middle|\gamma^{(\mathrm{t})}\right)\cong\frac{1}{\mathrm{M}}\sum_{\mathrm{k}=1}^{\mathrm{M}}\left\{\sum_{\mathrm{i}=1}^{\mathrm{N}}\log\mathrm{p}\left(\mathrm{y}_{\mathrm{i}}\middle|\mathbf{x}_{\mathrm{obs,i}},\mathbf{v}_{\mathrm{ik}},\mathrm{z}_{\mathrm{mis,i}},\beta\right)\right\}$$

$$\widehat{\mathrm{Q}}^{(2)}\left(\mu_{\mathrm{m}},\pi_{\mathrm{m}}\middle|\gamma^{(\mathrm{t})}\right)\cong\frac{1}{\mathrm{M}}\sum_{\mathrm{k}=1}^{\mathrm{M}}\left\{\sum_{\mathrm{i}=1}^{\mathrm{N}}\log\mathrm{p}\left(\mathbf{x}_{\mathrm{i}},\mathbf{z}_{\mathrm{i}}\middle|\mathbf{x}_{\mathrm{obs,i}},\mathrm{z}_{\mathrm{obs,i}},\mathbf{v}_{\mathrm{ik}},\mu_{\mathrm{m}},\pi_{\mathrm{m}}\right)\right\}$$

在 EM 方法中第 t 次遞迴的 M 步驟，即是計算使$\widehat{\mathrm{Q}}^{(1)}\left(\beta\middle|\gamma^{(\mathrm{t})}\right)$與$\widehat{\mathrm{Q}}^{(2)}\left(\mu_{\mathrm{m}},\pi_{\mathrm{m}}\middle|\gamma^{(\mathrm{t})}\right)$為最大值的$\beta^{(\mathrm{t+1})}$、$\mu_{\mathrm{m}}{}^{(\mathrm{t+1})}$與$\pi_{\mathrm{m}}{}^{(\mathrm{t+1})}$，由於感興趣的參數為$\beta^{(\mathrm{t+1})}$，因此只針對β做參數估計的收斂；在計算上，我們使用牛頓法(Newton Raphson；N-R)方法來計算$\beta^{(\mathrm{t+1})}$，即是利用：

$$\beta^{(\mathrm{t+1})}=\beta^{(\mathrm{t})}-\left[\mathrm{H}\left(\beta^{(\mathrm{t})}\right)\right]^{-1}\frac{1}{\partial\beta^{(\mathrm{t})}}\widehat{\mathrm{Q}}^{(1)}\left(\beta^{(\mathrm{t})}\middle|\gamma^{(\mathrm{t})}\right)\qquad(3.10)$$

其中

$$\frac{1}{\partial\beta^{(t)}}\widehat{Q}^{(1)}\left(\beta^{(t)}\middle|\,\gamma^{(t)}\right)=\begin{pmatrix}\frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{N}y_i\mathbf{x}_{i1}^{(k)}-b'(\gamma_1)\\ \frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{N}y_i\mathbf{x}_{i2}^{(k)}-b'(\gamma_2)\\ \vdots\\ \frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{N}y_i\mathbf{x}_{ip}^{(k)}-b'(\gamma_n)\end{pmatrix}。$$

且 Hessian matrix $H\left(\beta^{(t)}\right)=\frac{1}{\partial\beta^{(t)}\,\partial\beta^{(t)}}\widehat{Q}^{(1)}\left(\beta^{(t)}\middle|\,\gamma^{(t)}\right)$。

若根據先前對反應變項y_i的分佈設立，則$b'(\gamma_i)$為此函數$\frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{N}\log\left(1+\exp\left\{\eta_i^{(k)}\right\}\right)$對$\beta$做一階微分。透過(3.10)式的迭代，我們定義當第(t)次到第$(t+1)$遞迴所得到的參數估計之差值小於一特定值ε時，則此時的$\beta^{(t+1)}$為所求，即$\left\|\beta^{(t+1)}-\beta^{(t)}\right\|<\varepsilon$時則 N-R 法達到收斂；若尚未滿足收斂條件，則重複再對$\left[\mathbf{x}_{det,i}\middle|,y_i,\mathbf{x}_{obs,i},\mathbf{z}_{obs,i},\gamma^{(t)}\right]\times I\left(c_l<\mathbf{x}_{det,i}<c_u\right)$與$\left[\mathbf{z}_{mis,i}\middle|,y_i,\mathbf{x}_{obs,i},\mathbf{z}_{obs,i},\gamma^{(t)}\right]\times I\left(c_l<\mathbf{x}_{det,i}<c_u\right)$做取另一組抽樣樣本，如此循環後直至滿足收斂條件。

第四章　模擬研究

第一節　資料生成與模擬

假設一組資料$\{(y_i, \mathbf{x}_i^T, z_i); i = 1,2,\cdots\}$，$y_i = 0,1$為二元資料，$\mathbf{x}_i^T = (x_{i1}, x_{i2}, x_{i3})$為連續變數且皆受偵測極限影響；$z_i$為含有兩個項目(levels)的類別變項且具有缺失值，其遺失機制為隨機遺失機制。設立一線性模型為：

$$\text{logit}\, p(y_i = 1|\mathbf{x}_i, z_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 z_i \qquad (4.1)$$

則$p(y_i = 1|\mathbf{x}_i, z_i) = \exp\{\eta_i - \log(1 + e^{\eta_i})\}$，其中$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 z_i$，$(\beta_1, \beta_2, \beta_3, \beta_4) = (3, 0.26, 0.17, -0.5)$。其中$(x_{i1}, x_{i2}, x_{i3})$之偵測極限$(c_l, c_u)$的設立是根據其資料設限的比例而設，在這裡探討的設限比例為 10%、30%與 50%，因此偵測極限的上下限為使資料的對稱設限比例分別為 5%、15%與 25%之最大和最小值。

由於類別變數z_i的數值會影響連續變數$\mathbf{x}_i^T$的參數設立，因此先產生一組z_i值為$(0,1)$的資料且其機率為$p(z_i = m) = \pi_m m = 0,1$，並設立$(\pi_0, \pi_1) = (0.3, 0.7)$，則$z_i$值的產生方式如下：

(1)隨機生成一筆U_i，U_i~Uniform(0,1)。

(2)$z_i = \begin{cases} 0 & \text{若 } 0 \le U_i < 0.3 \\ 1 & \text{若 } 0.3 \le U_i < 1 \end{cases}$

重複上述(1)~(2)步驟，直至產生一組z_i的數值，$i = 1, \cdots, 300$。則(x_{i1}, x_{i2}, x_{i3})條件於$z_i = m$下的條件分佈服從多變量常態分佈$N_3(\mu_{im}, \Sigma_{3\times 3})$，期望值為$\mu_{im}$指的是類別變項的每一個項目所對應的$(x_{i1}, x_{i2}, x_{i3})$多變量常態分布之期望值；其設立為

$\mu_{im}^T = \begin{cases} [0,0,0]，當 z_i = 0 \\ [1,1,1]，當 z_i = 1 \end{cases}$。並令其共變異數矩陣(covariance matrix)為

$$\Sigma_{3\times 3} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix} = \begin{pmatrix} 1 & 0.3 & 0.4 \\ 0.3 & 1 & 0.7 \\ 0.4 & 0.7 & 1 \end{pmatrix};$$

藉由(4.1)式的模式設立，可生成300筆的y_i數值，其中y_i為服從一伯努力分佈$\mathrm{Ber}(p_i)$，且參數$p_i = p(y_i = 1|\mathbf{x}_i) = \exp\{\eta_i - \log(1 + e^{\eta_i})\}$。由於$(x_{i1}, x_{i2}, x_{i3})$受偵測極限$(c_l, c_u)$影響，故令落於偵測極限$(c_l, c_u)$外的數值為遺失資料。而具有隨機遺失資料的類別變數$z_i$之設定可透過先計算一遺失機率$p.\mathrm{mis}_i = \frac{\exp\{2\times y_i\}}{1+\exp\{2\times y_i\}}$，令一遺失指標變數$M_i$服從伯努力分佈$\mathrm{Ber}(p.\mathrm{mis}_i)$，則$z_i$的遺失資料設定為$z_i^* = \begin{cases} z_i \text{ 當 } M_i = 0 \\ \mathrm{NA} \text{ 當 } M_i = 1 \end{cases}$

根據條件機率的定義，第 i 個體的聯合機率$p(y_i, x_{i1}, x_{i2}, x_{i3}, z_i, \mu_m)$為：

$$
\begin{aligned}
&p(y_i, x_{i1}, x_{i2}, x_{i3}, z_i, \mu_m) \\
&= p(y_i|x_{i1}, x_{i2}, x_{i3}, z_i, \mu_m)p(x_{i1}, x_{i2}, x_{i3}|z_i, \mu_m)pr(z_i = m) \\
&= \left(\frac{\exp\{\eta_i\}}{1+\exp\{\eta_i\}}\right)^{I\{y_i=1\}}\left(\frac{1}{1+\exp\{\eta_i\}}\right)^{I\{y_i=0\}} \times \\
&\qquad \frac{1}{(2\pi)\sqrt{|\Sigma|}}\exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_m)\Sigma^{-1}(\mathbf{x}_i - \mu_m)^T\right\}\prod_{m=0}^{2}\pi_m^{I\{z_i=m\}} \qquad (4.2)
\end{aligned}
$$

對 (4.2)式取對數，則可得第 i 個體完整數據資料的對數概似函數：

$$
\begin{aligned}
&\ell_i(\beta, \mu_m, \pi_m|y_i, \mathbf{x}_i, z_i) = \ell_i(\gamma|y_i, \mathbf{x}_i, z_i) \\
&\qquad = \log[p(y_i|\mathbf{x}_i, z_i, \beta)] + \log[p(\mathbf{x}_i|z_i, \mu_m)] + \log[p(z_i|\pi_m)] \\
&\propto y_i\eta_i - \log(1 + e^{\eta_i}) - (\mathbf{x}_i - \mu_m)\sum^{-1}(\mathbf{x}_i - \mu_m)^T + \sum_{m=0}^{2} I\{z_i = m\}\log\pi_m \qquad (4.3)
\end{aligned}
$$

令$\gamma = (\beta, \mu_m, \pi_m)$。EM 法 E 步驟即對完整數據資料的對數概似函數取條件期望值，則 EM 法之第 i 個體的第$(t + 1)$次遞迴 E 步驟為

$$
\begin{aligned}
&Q_i(\gamma|\gamma^{(t)}) = E\left[\ell_i(\beta, \mu_m, \pi_m|y_i, \mathbf{x}_i, z_i)\middle|y_i, \mathbf{x}_{obs,i}, c_l < \mathbf{x}_{det,i} < c_u, z_{obs,i}, \gamma^{(t)}\right] \\
&= E[\log[p(y_i|\mathbf{x}_i, z_i, \beta)] + \log[p(\mathbf{x}_i|z_i, \mu_m)] \\
&\qquad + \log[p(z_i|\pi_m)]\big|y_i, \mathbf{x}_{obs,i}, c_l < \mathbf{x}_{det,i} < c_u, z_{obs,i}, \gamma^{(t)}\big] \\
&= \sum_{z_{mis,i}}\int \log[p(y_i|\mathbf{x}_i, z_i, \beta)] \times p\left(x_{det,i}, z_{mis,i}\middle|y_i, \mathbf{x}_{obs,i}, z_{obs,i}, \gamma^{(t)}\right) \\
&\qquad \times I\left(c_l < \mathbf{x}_{det,i} < c_u\right)dx_{det,i} \\
&+ \sum_{z_{mis,i}}\int\{\log[p(\mathbf{x}_i|z_i, \mu_m)] + \log[p(z_i|\pi_m)]\} \\
&\qquad \times p\left(x_{det,i}, z_{mis,i}\middle|y_i, \mathbf{x}_{obs,i}, z_{obs,i}, \gamma^{(t)}\right)I\left(c_l < \mathbf{x}_{det,i} < c_u\right)dx_{det,i} \qquad (4.4)
\end{aligned}
$$

其中參數$\gamma^{(t)} = \left(\beta^{(t)}, \mu_m^{(t)}, \pi_m^{(t)}\right)$。在(4.4)式中，隨著$x_{det,i}$所包含的變數個數增多時，其積分計算將隨之變難，即若假設第 i 個體受偵測極限影響而有遺失值的$x_{det,i} = (x_{i1}, x_{i2}, x_{i3})$，則(4.4)的積分式為三維積分。因此我們使用蒙地卡羅積分來解決取代多重積分的問題，即 E 步驟的多重積分我們需透過對$\left[\mathbf{x}_{det,i}, \mathbf{z}_{mis,i}\middle|, y_i, \mathbf{x}_{obs,i}, \mathbf{z}_{obs,i}, \gamma^{(t)}\right] \times I(c_l < \mathbf{x}_{det,i} < c_u)$分佈抽取遺失資料$(\mathbf{x}_{det,i}, \mathbf{z}_{mis,i})$，而此抽樣的步驟我們可利用 Gibbs 抽樣方法對滿條件分佈$\left[\mathbf{x}_{det,i}\middle|, y_i, \mathbf{x}_{obs,i}, \mathbf{z}_{obs,i}, \gamma^{(t)}\right] \times I(c_l < \mathbf{x}_{det,i} < c_u)$與$\left[\mathbf{z}_{mis,i}\middle|, y_i, \mathbf{x}_{obs,i}, \mathbf{z}_{obs,i}, \gamma^{(t)}\right] \times I(c_l < \mathbf{x}_{det,i} < ub)$做抽樣，抽取 1000 個隨機樣本向量$\mathbf{v}_i$，$\mathbf{v}_i$為第 i 個個體中具有遺失資料的變數，並捨棄(burn-in)前 750 個樣本，根據蒙地卡羅積分的定義，EM 法的 E 步驟會近似為：

$$E\left[\log[p(y_i|\mathbf{x}_i, z_i, \beta)]\middle|y_i, \mathbf{x}_{obs,i}, c_l < \mathbf{x}_{det,i} < c_u, z_{obs,i}, \gamma^{(t)}\right]$$

$$\cong \frac{1}{250}\sum_{k=1}^{250}\left(\log\left[p\left(y_i\middle|\mathbf{x}_{obs,i}, \mathbf{v}_{ik}, z_{mis,i}, \beta\right)\right]\right) = Q_i^{(1)}\left(\beta\middle|\gamma^{(t)}\right) \quad (4.5)$$

$$E\left[\log[p(\mathbf{x}_i|z_i, \mu_m)] + \log[p(z_i|\pi_m)]\middle|y_i, \mathbf{x}_{obs,i}, c_l < \mathbf{x}_{det,i} < c_u, z_{obs,i}, \gamma^{(t)}\right]$$

$$\cong \frac{1}{250}\sum_{k=1}^{250}\log\left[p\left(\mathbf{x}_i, z_i\middle|\mathbf{x}_{obs,i}, z_{obs,i}, \mathbf{v}_{ik}, \mu_m, \pi_m\right)\right] = Q_i^{(2)}\left(\mu_m, \pi_m\middle|\gamma^{(t)}\right) \quad (4.6)$$

為了能從滿條件分佈抽樣，我們可知道：

$$\left[\mathbf{x}_{det,i}\middle|, y_i, \mathbf{x}_{obs,i}, \mathbf{z}_{mis,i}, \mathbf{z}_{obs,i}, \gamma^{(t)}\right] I(c_l < \mathbf{x}_{det,i} < c_u) \propto \left[y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}\right] \times \left[\mathbf{x}_{mis,i}\middle|\mathbf{x}_{obs,i}, \mathbf{z}_i\right] I(c_l < \mathbf{x}_{det,i} < c_u) \quad (4.7)$$

$$\left[\mathbf{z}_{mis,i}\middle|, y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{obs,i}, \gamma^{(t)}\right] I(c_l < \mathbf{x}_{det,i} < c_u) \propto \left[y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}\right] \times \left[\mathbf{x}_{mis,i}\middle|\mathbf{x}_{obs,i}, \mathbf{z}_i\right] \times \left[\mathbf{z}_{mis,i}\right] I(c_l < \mathbf{x}_{det,i} < c_u) \quad (4.8)$$

其中，

$\left[y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}\right]$的機率密度函數為$\left(\frac{\exp\{\eta_i\}}{1+\exp\{\eta_i\}}\right)^{I\{y_i=1\}}\left(\frac{1}{1+\exp\{\eta_i\}}\right)^{I\{y_i=0\}}$

$\left[\mathbf{x}_{mis,i}\middle|\mathbf{x}_{obs,i}, \mathbf{z}_i\right]$的機率密度函數可依據$\mathbf{x}_{mis,i}$的變數個數不同而不同。在此模擬中，令$\mathbf{x}_{mis,i}$與$\mathbf{x}_{obs,i}$所對應的期望值分別為$\mu_{mis,i}$和$\mu_{obs,i}$，共變異數矩陣又可劃分成

$\Sigma = \begin{pmatrix} \sum_{mis,mis} & \sum_{mis,obs} \\ \sum_{obs,mis} & \sum_{obs,obs} \end{pmatrix}$，則$[\mathbf{x}_{mis,i}|\mathbf{x}_{obs,i}, \mathbf{z}_i] \sim N(\mu_{mis,i} + \sum_{mis,obs}\sum_{obs,obs}^{-1}(\mathbf{x}_{obs,i} - \mu_{obs,i}), \sum_{mis,mis} - \sum_{mis,obs}\sum_{obs,obs}^{-1}\sum_{obs,mis})$

$[\mathbf{z}_{mis,i}]$的機率密度函數為$\pi_0^{I\{z_i=0\}}\pi_1^{I\{z_i=1\}}$

由於滿條件分佈又可寫成(4.7)式與(4.8)式的右側型式，因此我們可使用 ARMS 抽樣方法(Gilks, Best, & Ta,1995)來抽取滿條件分佈。抽樣後，由(4.5)與(4.6)式的可知$Q^{(1)}(\beta|\gamma^{(t)})$與$Q^{(2)}(\mu_m, \pi_m|\gamma^{(t)})$函數可表示成

$$\widehat{Q}^{(1)}(\beta|\gamma^{(t)}) \cong \sum_{i=1}^{300} Q_i^{(1)}(\gamma|\gamma^{(t)}) = \frac{1}{250}\sum_{k=1}^{250}\left\{\sum_{i=1}^{300} y_i\eta_i^{(k)} - \sum_{i=1}^{200}\log\left(1+e^{\eta_i^{(k)}}\right)\right\} \quad (4.9)$$

$$\widehat{Q}^{(2)}(\mu_m, \pi_m|\gamma^{(t)}) \cong \sum_{i=1}^{300} Q_i^{(2)}(\gamma|\gamma^{(t)})$$

$$\cong \frac{1}{250}\sum_{k=1}^{250}\{\sum_{i=1}^{300}(\mathbf{x}_i - \mu_m)\Sigma^{-1}(\mathbf{x}_i - \mu_m)^T + \sum_{m=0}^{1}(\sum_{i=1}^{300} W_{im})\pi_m\} \quad (4.10)$$

其中$\eta_i^{(k)} = \beta_1 x_{i1}^{(k)} + \beta_2 x_{i2}^{(k)} + \beta_3 x_{i3}^{(k)} + \beta_4 z_i^{(k)}$。

而 EM 演算法的 M 步驟為求使得$\ell(y_i, \mathbf{x}_i, z_i|\beta, \mu_m, \pi_m)$有最大值的參數$\hat{\gamma} = (\beta, \mu_m, \pi_m)$，亦即求$\hat{\gamma}^{(t)} = \left(\beta^{(t+1)}, \mu_m^{(t+1)}, \pi_m^{(t+1)}\right) = \arg\max Q(\gamma|\gamma^{(t)})$。

我們利用 Newton-Raphson 方法估計$\beta^{(t+1)}$：

$$\beta^{(t+1)} = \beta^{(t)} - \left[H(\beta^{(t)})\right]^{-1}\frac{1}{\partial\beta^{(t)}}\widehat{Q}^{(1)}(\beta^{(t)}|\gamma^{(t)})$$

$$\frac{1}{\partial\beta^{(t)}}\widehat{Q}^{(1)}(\beta^{(t)}|\gamma^{(t)}) = \begin{pmatrix} \frac{1}{250}\sum_{k=1}^{250}\left(\sum_{i=1}^{300} y_i x_{i1}^{(k)}\right) - \frac{1}{250}\sum_{k=1}^{250}\sum_{i=1}^{300}\left(\frac{x_{i1}^{(k)}e^{\eta_i^{(k)}}}{1+e^{\eta_i^{(k)}}}\right) \\ \frac{1}{250}\sum_{k=1}^{250}\left(\sum_{i=1}^{300} y_i x_{i2}^{(k)}\right) - \frac{1}{250}\sum_{k=1}^{250}\sum_{i=1}^{300}\left(\frac{x_{i2}^{(k)}e^{\eta_i^{(k)}}}{1+e^{\eta_i^{(k)}}}\right) \\ \frac{1}{250}\sum_{k=1}^{250}\left(\sum_{i=1}^{300} y_i x_{i3}^{(k)}\right) - \frac{1}{250}\sum_{k=1}^{250}\sum_{i=1}^{300}\left(\frac{x_{i3}^{(k)}e^{\eta_i^{(k)}}}{1+e^{\eta_i^{(k)}}}\right) \\ \frac{1}{250}\sum_{k=1}^{250}\left(\sum_{i=1}^{300} y_i z_i^{(k)}\right) - \frac{1}{250}\sum_{k=1}^{250}\sum_{i=1}^{300}\left(\frac{z_i^{(k)}e^{\eta_i^{(k)}}}{1+e^{\eta_i^{(k)}}}\right) \end{pmatrix} \quad (4.11)$$

$$\text{Hessian matrix } H(\beta^{(t)}) = \frac{1}{\partial\beta^{(t)}\,\partial\beta^{(t)}}\widehat{Q}^{(1)}(\beta^{(t)}|\gamma^{(t)}) \quad (4.12)$$

對於$\widehat{Q}^{(2)}(\mu_m, \pi_m|\gamma^{(t)})$中參數$\mu_m$的估計為：

$$\hat{\mu}_m = \left(\frac{1}{250}\sum_{k=1}^{250}\left(\sum_{i=1}^{300} x_{i1}^{(k)}\right), \frac{1}{250}\sum_{k=1}^{250}\left(\sum_{i=1}^{300} x_{i2}^{(k)}\right), \frac{1}{250}\sum_{k=1}^{250}\left(\sum_{i=1}^{300} x_{i3}^{(k)}\right)\right) \quad (4.13)$$

$$\hat{\pi}_m = \frac{1}{250}\sum_{k=1}^{250}\left(\sum_{i=1}^{300} z_i^{(k)}\right) \quad (4.14)$$

$m = 0,1$。在 10%、30%與 50%的設限比例下，當第(t)次到第$(t+1)$遞迴所得到的參數估計之差值分別小於0.01、0.1 與 1 時，則此時的$\beta^{(t+1)}$為所求，即 $\left\|\beta^{(t+1)} - \beta^{(t)}\right\| < \varepsilon$時則 N-R 法達到收斂，其中$\varepsilon = 0.01$、0.1 與 1；若尚未滿足收斂條件，則重複再對$\left[\mathbf{x}_{det,i}\middle|, y_i, \mathbf{x}_{obs,i}, \mathbf{z}_{obs,i}, \gamma^{(t)}\right] \times I\left(c_l < \mathbf{x}_{det,i} < c_u\right)$與 $\left[\mathbf{z}_{mis,i}\middle|, y_i, \mathbf{x}_{obs,i}, \mathbf{z}_{obs,i}, \gamma^{(t)}\right] \times I\left(c_l < \mathbf{x}_{det,i} < c_u\right)$做取另一組抽樣樣本，並計算(4.5)式至(4.14)式，如此循環後直至滿足收斂條件。

第二節　模擬結果

我們以 Schisterman *et al.*(2006)參數估計之做法和完整觀察個體分析(Complete-Case Analysis)與本研究所提的方法作參數估計的比較，進行模擬 250 組資料的參數估計。表 1、表 2 與表 3 為此三種方法分別在資料設限比例為 10%、30%與 50%的情形下，設立真實迴歸參數值$(\beta_1, \beta_2, \beta_3, \beta_4) = (3, 0.26, 0.17, -0.5)$作 250 組模擬資料之參數估計結果與標準差的比較。

從表 1 可知道，在設限比例為 10%下，以 Schisterman *et al.*(2006)與本文所提之方法在參數估計上較完整觀察個體分析來得不偏，就精準度與不偏性來說，又以本文的方法較為精準與不偏；在比較中，三種方法在估計類別變數的參數上都有較明顯的偏差，且其偏差範圍皆大於 0.1。

就設限比例為 30%的情況下，從表 2 可得知三種估計方法與真實參數值有些許明顯的差異，由於在分析上，完整觀察個體分析刪除多筆資料，因此在估計上其所估計之參數較偏離真實直，因此在三種方法的比較中，尤以完整觀察個體分析的參數估計與真實參數值偏差為最大；而本文的方法在設限比例變大下，在估計參數上也較明顯有偏差，但仍為三種比較方法中為最佳的估計方法。

表 3 中，當連續變數的設限比例增加到 50%時，三種估計方法在參數估計上有明顯的偏差，且以完整觀察個體分析之參數估計上的偏差為最大，而本文之方法與真實參數值之差異也隨設限比例之增加而隨之變大；由於本文之方法因探討到變數間彼此的相關性，因此在三者比較中，本文之方法較 Schisterman *et al.* (2006)與完整觀察個體分析來的較佳，以完整觀察個體分析為較差。

表 1：設限比例為 10%下，Monte Carlo EM、Schisterman *et al.*(2006)與 Complete-Case Analysis 之比較

		Monte Carlo EM		Schisterman *et al.*(2006)		Complete-Case Analysis	
	True Value	Estimate	SE	Estimate	SE	Estimate	SE
β_1	3	2.99752	0.01044	3.07564	0.03337	3.16603	0.17750
β_2	0.26	0.25987	0.01275	0.24142	0.01815	0.44582	0.16052
β_3	0.17	0.17063	0.01727	0.31409	0.01935	0.46533	0.29617
β_4	-0.5	-0.50014	0.00364	-0.82545	0.02279	-0.43484	0.27418
μ_{01}	0	0.08661	0.00772	0.09375	0.00488	0.31450	0.02676
μ_{02}	0	0.00298	0.00193	0.67756	0.00574	0.74547	0.03146
μ_{03}	0	0.00175	0.01889	0.93887	0.00473	0.10316	0.08089
μ_{11}	1	0.99172	0.00394	0.93750	0.00488	0.31423	0.03392
μ_{12}	1	1.00693	0.01090	0.67356	0.00574	0.69451	0.03857
μ_{13}	1	1.02047	0.01002	0.91888	0.00473	1.02034	0.03764
π_1	0.3	0.40105	0.00745	0.53420	0.02741	0.48451	0.02179

*註：$(\mu_{01},\mu_{02},\mu_{03})$代表類別變數之數$z$值為 0 時，連續變數$(x_1,x_2,x_3)$分佈所對應的期望值；$(\mu_{11},\mu_{12},\mu_{13})$代表類別變數之數$z$值為 1 時，連續變數$(x_1,x_2,x_3)$分佈所對應的期望值；$\pi_1$為類別變數$z$之數值為 0 的機率。

表 2：設限比例為 30%下，Monte Carlo EM、Schisterman *et al.*(2006)與 Complete-Case Analysis 之比較

		Monte Carlo EM		Schisterman *et al.*(2006)		Complete-Case Analysis	
	True Value	Estimate	SE	Estimate	SE	Estimate	SE
β_1	3	2.79837	0.01128	2.29682	0.03036	3.30735	0.03775
β_2	0.26	0.18385	0.01899	0.11072	0.01982	0.35425	0.02136
β_3	0.17	0.30872	0.01171	0.31739	0.02053	0.44315	0.02161
β_4	-0.5	-0.78020	0.00199	-1.14014	0.02493	-0.18828	0.02294
μ_{01}	0	0.18468	0.00960	0.31932	0.04632	0.33742	0.00515
μ_{02}	0	0.32539	0.00821	0.62226	0.00545	0.59464	0.00559
μ_{03}	0	0.51390	0.01261	0.75200	0.00504	0.82250	0.00531
μ_{11}	1	0.77893	0.00146	0.31932	0.04632	0.34116	0.00374
μ_{12}	1	0.82206	0.00180	0.62226	0.00545	0.60567	0.00407
μ_{13}	1	1.21822	0.04373	0.85200	0.00504	0.84019	0.00390
π_1	0.3	0.35886	0.00152	0.50432	0.02648	0.48382	0.02153

*註：$(\mu_{01},\mu_{02},\mu_{03})$代表類別變數之數Z值為 0 時，連續變數$(x_1,x_2,x_3)$分佈所對應的期望值；$(\mu_{11},\mu_{12},\mu_{13})$代表類別變數之數Z值為 1 時，連續變數$(x_1,x_2,x_3)$分佈所對應的期望值；$\pi_1$為類別變數Z之數值為 0 的機率。

表 3：設限比例為 50%下，Monte Carlo EM、Schisterman *et al.*(2006)與 Complete-Case Analysis 之比較

		Monte Carlo EM		Schisterman *et al.*(2006)		Complete-Case Analysis	
	True Value	Estimate	SE	Estimate	SE	Estimate	SE
β_1	3	2.59837	0.01520	1.39703	0.02824	6.25357	1.53243
β_2	0.26	0.14385	0.09179	-0.15641	0.02538	-0.25712	1.04256
β_3	0.17	0.36872	0.02103	0.03511	0.02523	2.37343	1.71283
β_4	-0.5	-0.98525	0.02031	-1.67454	0.02069	-1.13779	0.54129
μ_{01}	0	0.37832	0.00991	0.46202	0.00392	0.43160	0.01580
μ_{02}	0	0.43376	0.08370	0.59096	0.00418	0.58071	0.06358
μ_{03}	0	0.60102	0.01329	0.64377	0.00401	0.72735	0.03601
μ_{11}	1	0.61353	0.01815	0.46202	0.00392	0.44594	0.03948
μ_{12}	1	0.66830	0.01813	0.59096	0.00418	0.58500	0.00453
μ_{13}	1	1.54920	0.04641	0.74377	0.00401	0.62828	0.01416
π_1	0.3	0.49373	0.01699	0.50954	0.02605	0.52336	0.02114

*註：$(\mu_{01}, \mu_{02}, \mu_{03})$代表類別變數之數z值為 0 時，連續變數$(x_1, x_2, x_3)$分佈所對應的期望值；$(\mu_{11}, \mu_{12}, \mu_{13})$代表類別變數之數z值為 1 時，連續變數$(x_1, x_2, x_3)$分佈所對應的期望值；$\pi_1$為類別變數z之數值為 0 的機率。

第五章　結果與討論

不同於 May, Ibrahim, & Chu (2011)研究方法僅能適用在變數為類別變數下以 Monte Carlo EM 法進行迴歸參數的估計，並與完整觀察個體分析和簡單替代法比較，本篇論文在模式中額外加入類別變數，並探討當類別變數為隨機遺失機制且連續變數為受偵測極限影響的非隨機遺失值機制時，藉由 Monte Carlo EM 演算法計算其參數估計值，並與 Schisterman *et al.*(2006)方法和完整觀察個體分析在不同設限比例下作參數估計值的比較。

經由比較在不同設限比例下三種方法的參數估計值可知，由於因為本篇論文考慮到變數間彼此的相關性，故所提出的方法在不同的設限比例下其參數估計值較 Schisterman *et al.*(2006)方法和完整觀察個體分析來得不偏。然而隨著設限比例的增加，本研究方法的參數估計值與真實係數值也隨之有明顯的差距，但仍較其他另兩種方法來得接近真實值。因此，本文提供一個相較於個別插補和完全觀察個體分析其參數估計較為不偏的方法。

由於本文所指的設限比例只針對遺失的連續變數資料占總資料的比例而言，並未考慮到類別變數的遺失比例，而導致類別變數的遺失比例過多，進而影響到本論文方法的估計精準度。因此，在未來研究時，可進一步將類別的遺失比例也一併考慮；另外，也可探討當隨機遺失機制的類別變數為非隨機遺失機制時，此時的參數估計是否仍為三種比較方法中的最佳估計量，且與隨機遺失機制的參數估計量是否有差異。這些議題皆可放於日後作為探討的目標。

參考文獻

1. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological, 39*(1), 1-38.
2. D'Anaelo, G., Weissfeld, L., & Investigators, G. (2008). An index approach for the Cox model with left censored covariates. *Statistics in Medicine, 27*(22), 4502-4514. doi: Doi 10.1002/Sim.3285
3. Ibrahim, J. G. (1990). Incomplete Data in Generalized Linear-Models. *Journal of the American Statistical Association, 85*(411), 765-769.
4. Lipsitz, S. R., & Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika, 83*(4), 916-922.
5. Little, R. J. A., & Schluchter, M. D. (1985). Maximum-Likelihood Estimation for Mixed Continuous and Categorical-Data with Missing Values. *Biometrika, 72*(3), 497-512
6. Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley
7. Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., . . . Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives, 112*(17), 1691-1696. doi: Doi 10.1289/Ehp.7199
8. Lynn, H. S. (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine, 20*(1), 33-45
9. May, R. C., Ibrahim, J. G., & Chu, H. T. (2011). Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statistics in Medicine, 30*(20), 2551-2561. doi: Doi 10.1002/Sim.4280
10. MacDougall, Daniel, Crummett, Warren B., *et al* (1980). "Guidelines for Data Acquisition and Data Quality Evaluation in Environmental Chemistry." Analytical Chemistry **52**(14): 2242-2249
11. Nie, L., Chu, H. T., Liu, C. L., Cole, S. R., Vexler, A., & Schisterman, E. F. (2010). Linear Regression With an Independent Variable Subject to a Detection Limit. *Epidemiology, 21*, S17-S24. doi: Doi 10.1097/Ede.0b013e3181ce97d8
12. Richardson, D. B., & Ciampi, A. (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology, 157*(4), 355-363. doi: Doi 10.1093/Aje/Kwf217
13. Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*(3), 581-590

14. Rigobon, R., & Stoker, T. M. (2009). Bias From Censored Regressors. *Journal of Business & Economic Statistics, 27*(3), 340-353. doi: DOI 10.1198/jbes.2009.06119
15. Schisterman, E. F., Vexler, A., Whitcomb, B. W., & Liu, A. Y. (2006). The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology, 163*(4), 374-383. doi: Doi 10.1093/Aje/Kwj039
16. Stubbendick, A. L., & Ibrahim, J. G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica, 16*(4), 1143-1167.
17. Wei, G. C. G., & Tanner, M. A. (1990). A Monte-Carlo Implementation of the Em Algorithm and the Poor Mans Data Augmentation Algorithms. *Journal of the American Statistical Association, 85*(411), 699-704

【附錄】

(3.8)式與(3.9)的推導過程如下：

(3.8)式：$p(\mathbf{x}_{det,i}|y_i, \mathbf{x}_{obs,i}, \mathbf{z}_{mis,i}, \mathbf{z}_{obs,i}, \gamma^{(t)})I(c_l < \mathbf{x}_{det,i} < c_u)$

$$= \frac{p(y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{mis,i}, \mathbf{z}_{obs,i}, \gamma^{(t)})I(c_l < \mathbf{x}_{det,i} < c_u)}{\int p(y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{mis,i}, \mathbf{z}_{obs,i}, \gamma^{(t)})d\mathbf{x}_{det,i}}$$

$$\propto p(y_i, \mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})I(c_l < \mathbf{x}_{det,i} < ub)$$

$$= p(y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}) \times p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})\, I(c_l < \mathbf{x}_{det,i} < c_u)$$

$$= p(y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}) \times \frac{p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})}{\iint p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})d\mathbf{x}_{obs,i}d\mathbf{x}_{det,i}} \times \iint p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})d\mathbf{x}_{obs,i}d\mathbf{x}_{det,i} \times I(c_l < \mathbf{x}_{det,i} < c_u)$$

$$= p(y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}) \times p(\mathbf{x}_i|\mathbf{z}_i, \gamma^{(t)}) \times \iint p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})d\mathbf{x}_{obs,i}d\mathbf{x}_{det,i} \times I(c_l < \mathbf{x}_{det,i} < c_u)$$

$$\propto\ p(y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}) \times p(\mathbf{x}_i|\mathbf{z}_i, \gamma^{(t)})$$

(3.9)式：$p(\mathbf{z}_{mis,i}|y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{obs,i}, \gamma^{(t)})I(c_l < \mathbf{x}_{det,i} < c_u)$

$$= \frac{p(y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{mis,i}, \mathbf{z}_{obs,i}, \gamma^{(t)})I(c_l < \mathbf{x}_{det,i} < c_u)}{\sum_{\mathbf{z}_{mis,i}} p(y_i, \mathbf{x}_{obs,i}, \mathbf{x}_{det,i}, \mathbf{z}_{mis,i}, \mathbf{z}_{obs,i}, \gamma^{(t)})}$$

$$\propto p(y_i, \mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})I(c_l < \mathbf{x}_{det,i} < c_u)$$

$$= p(y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}) \times p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})\, I(c_l < \mathbf{x}_{det,i} < c_u)$$

$$= p(y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}) \times \frac{p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})}{\iint p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})d\mathbf{x}_{obs,i}d\mathbf{x}_{det,i}} \times \iint p(\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)})d\mathbf{x}_{obs,i}d\mathbf{x}_{det,i} \times I(c_l < \mathbf{x}_{det,i} < c_u)$$

$$= p(y_i|\mathbf{x}_i, \mathbf{z}_i, \gamma^{(t)}) \times p(\mathbf{x}_i|\mathbf{z}_i, \gamma^{(t)}) \times p(\mathbf{z}_i|\gamma^{(t)}) \times I(c_l < \mathbf{x}_{det,i} < c_u)$$