二、研究計畫內容（以 10 頁為限）：

(一)摘要

選模雖然只是在迴歸分析的一個小環節，但也是相當重要的角色，選模結果的好與壞更是會直接反映在研究的結果，所以選模的重要性絕對不容小覷，本計畫提出有別於以往自動選模(向前法、向後法、逐步迴歸法)及僅考慮「單一」準則的選模方式，計畫將會考慮模型的超數據(metadata)，接著利用群集分析為所有可能的模型分群，並從分群結果挑選出表現合宜的模型，且計畫的進程將全面使用 R 軟體，不僅提升效率，也可將計畫的成果提供給其他研究者另一種選模的方式及軟體工具。

(二)研究動機與研究問題

Konishi & Kitagawa (2008)說過：「大部分統計推論的問題都可以被認為涉及統計模型」[1]，其中，迴歸模型更是經常被使用的統計模型，一個好的迴歸模

型，可以有效率的實現研究目的。

然而迴歸模型的解釋變數越多，可能會使模型難以維護，另外較少的解釋變數會使模型較容易使用及了解[2]。根據前述理由，一般推薦變數較少的模型，所以選模的過程也應該更加的嚴謹、慎重，然而現在是資訊量爆炸的時代，資料集的變數持續增加，即使經過篩選仍然會留下很多，這個時候若要做迴歸分析勢必會對要選擇哪個模型為最佳的模型，有所疑慮，根據傳統的作法，向前法、向後法、逐步迴歸法，僅只能藉由模型的 F 值，讓程式自動為我們選擇最後的模型；另一種做法，則是僅利用「單一」準則，判斷所有模型的優劣，但這些作法，只考慮模型眾多資訊的其中一個，結果可能會有偏頗，而其他被捨棄的模型，可能存在著更適合用來預測目標變數的模型。

為了彌補傳統作法的缺點，本研究想做的是留下資料集所有可能的模型，並留下模型的超數據(metadata)：同時考慮「多種」選模準則，諸如 AIC、BIC、PRESS、Cp statistics、s^2、R^2_{adj}，**再透過 k-means 群集分析將模型分群(1 群以上)，如此將產生與群數個數相同的候選模型。如果只考慮一個準則、不分群(k=1)，就會得到傳統的選模結果。Joseph B. Kadan (2004)說過：選模是任何統計分析重要的組成成分，為了節省成本和時間，無法對所有模型做驗證[3]，所以經過群集分析，資料集只需將 k 個候選模型交叉驗證，挑選最適合該資料集的模型。**

(三)文獻回顧與探討

1. 選模(Model selection)：從一組候選模型的集合，評估不同模型的表現，從而選擇出一個最佳的統計模型，且選取模型的方法大致分為四類[4]：向前法(forward method)、向後法(backward method)、逐步迴歸法(stepwise regression method)；另外還有一種是所有可能的迴歸模型分別使

用 AIC、BIC、PRESS、Cp statistics [5]、s^2、R^2_{adj}等準則判斷。本計畫同時利用多種判斷準則為所有模型分群。

2. 多變量統計分析（Multivariate Statistical Analysis),又稱多元統計分析，為統計學的一支，常用於管理科學、社會科學和生命科學等領域。多變量分析主要用於分析擁有多個變數的資料，探討資料彼此之間的關聯性或是釐清資料的結構，而有別於傳統統計方法所著重的參數估計以及假設檢定。**由於多變量分析方法需要複雜且大量的計算，因此多藉助電腦運算。常見的分析方法：主成分分析、因素分析、群集分析[6]。本計畫會使用群集分析作為研究方法。**

3. 群集分析(Cluster Analysis)：是一種一般邏輯程序，它能根據相似性與相異性，客觀地將相似者歸集在同一群集內，群集分析的目的在辨認某些特性上相似的迴歸線，並把這些迴歸線按照這些特性劃分成幾個群集，使在同一群集內的迴歸線具有高度的同質性，而不同群集的迴歸線應彼此遠隔，群集分析並不是一種統計推論技術，而是將一組觀察值的結構特性數量化的一種客觀方法。相似性的衡量可分成：距離衡量(distance measures)和關聯衡量 (association measures) [7]；而常見群集分析的方法，可分為兩大類，4種基本方法[8]：

- 階層式群集方法（hierarchical methods）

(1)單一連結法（single linkage）

(2)質心法（centroid method）

(3)華德最小變異法(Ward's minimum variance method）

- 非階層式群集方法（non-hierarchical methods）

(4)K 平均數法（k-means methods）

階層式群集方法使用各迴歸線間距離或組內差異值比較兩條迴歸線最接近合併為同一群集內，合併後迴歸線間距離或組間內差異值再次比較，直到研究迴歸線都被合併到適合的群集內。

非階層式群集方法是將研究者已知（known）或預先假設群集個數 k，每個迴歸線會依照與每個群集 k 之質心（centroid）最小距離加以比較分配到最短距離之群集，每條迴歸線依照此標準（criterion）驗證直到完全到適合的群集內。最常使用為 K 平均數法（k-means methods）來分群[9]。

本計畫採用距離衡量模型的相似性，群集方法選擇非階層式的 K 平均數法。

4. 距離衡量：很多相似性的衡量是以點與點間的距離為代表，常見的有以下兩種類型[10]：

- 歐幾里得距離(Euclidean distance)：為兩點間最短的直線距離，是評估彼此相似程度最常見的測量方法。
- 曼哈頓距離(Manhattan distance)：為 L1-距離或城市區塊距離，也就是在歐幾里得空間的固定直角坐標系上兩點所形成的線段對軸產生的投影的距離總和。

本計畫採用「歐幾里得距離」衡量模型的相似性。

5. 超數據(Metadata)：是結構化的資訊，描述資料本身特性的資料，用來形容、解釋、定位或是使資料本身更加容易檢索、使用和管理的訊息資源，通常被稱為數據的數據，一般分為 3 種類型[11]：

- 描述型：目的在於發現和鑑定的數據，例如：標題、摘要、作者、關鍵字等。
- 結構型：指出複合資源如何集合在一起，例如：頁面和章節之間的組合順序。
- 管理型：提供管理資源的幫助訊息，例如：數據是何時或者如何創建、文件的類型、技術性的資訊、誰可以取得等等。

創建超數據的重要原因，是為了便於發現相關信息，除了資源的發現，超數據也可幫助組織電子資源，方便互操作性和傳統資源的整合，提供數字的識別、支持存檔和保存，被稱為確保資源能生存並繼續抵達未來的關鍵。本計畫會收集迴歸模型的超數據，提供分群之用。

6. K 次交叉驗證(k-fold cross-validation)：以取後不放回的方式，將數據集隨機分割成 k 個子集，各子集大小大約相同[12]。一個單獨的子樣本被保留作為驗證模型的數據，其他 k-1 個樣本用來訓練。交叉驗證重複 k 次，每個子樣本驗證一次，並且平均 k 次分群準確度的估計值。這個方法的優勢在於，同時重複運用隨機產生的子樣本進行訓練和驗證，10 次交叉驗證是最常用的[13]。

(四)研究方法及步驟

本研究使用開放原始碼的統計軟體 R，作為研究工具。然後結合建模及 k-means 群集分析作為研究方法。為使計畫撰寫更加順利，本研究以在多變量分析課程中取得的學習歷程數據以及從 UCI[14]取得的其他資料集為例子，進行整個計畫的撰寫及驗證。

- 第一步(數據前處理)：

首先針對資料集，先編寫程式收集數據的相關資訊以及必要的數據前處理作業，包括資料的維度、變數名稱等，以及重新編碼、合併、設定變數屬性等。

- 第二步(產生所有迴歸模型)：

接著利用 R 軟體的套件「leaps」，產生所有可能的迴歸模型，方法為「exhaustive」[15]，本研究分成三個階段，第一階段先考慮每個變數的納入與剔除(例如：若有 10 個解釋變數，則會有 2^{10}=1024 個模型)；第二階段則是將一階交

互作用項的解釋變數納入考慮；第三階段為考慮二階交互作用或二階以上的交互作用項，計畫預計以第一、二階段為主，如果時間允許，將會發展到第三階段。

- 第三步(收集模型的超數據)：

收集所有候選模型的超數據，計畫收集所有迴歸線判斷模型優劣的準則，諸如：AIC、BIC、PRESS、Cp statistics、s^2、R^2_{adj}的數值，以上僅列出六種數值，但計畫實際執行過程中，可能會設計其他的判斷準則。

- 第四步(群集分析)：

接著使用群集分析收集的超數據，這步驟分為上半部的距離衡量和下半部的 K 平均數法。上半部：將收集的超數據衡量模型的距離，採用 R 軟體計算距離的套件「cluster」[16]，計算「歐基里德距離」；下半部：使用同一套件，函數選擇 K 平均數法「kmeans」。另外 J.Hair (1995)說過「研究發現，如果群集數介於三個到六個之間，較易解釋每個群集的特徵」[17]，所以本研究的群集個數定為 3，如果實際情況不可行，會以相差距離為依據，另行決定群集個數，但至少會 2 群，方便研究與傳統作法(只考慮一個準則、不分群)的相異處。

- 第五步(挑選『好的模型』所屬的那群)：

延續上個步驟，觀察 3 個群集中的模型，以 BIC 較小為準則找出表現較好的模型群集。

- 第六步(實證)：

研究的最後一個步驟，使用上步驟取得的某一群集多個模型，利用 UCI 的數據集及多變量分析課程中所取得的學習歷程數據為例，使用 R 軟體的套件「cvTools」[18]，實際驗證多個模型的優劣，驗證方法採「K 次交叉驗證」， k 定為 10，並

且平均 10 次的結果，取得最終分群準確度的估計值；並比較多個估計值，得到最適合該資料集的最終模型。

(五)預期結果

就整個計畫的過程，本研究把預期結果共分為三類：計畫內容的成果、R 軟體的程式設計、個人能力的提升。

1. 計畫內容的成果：根據以上的研究步驟，預期可以得到一種利用群集分析充分考慮所有資訊的選模方式，其中記錄、儲存了各資料集的變數資訊、模型的超數據；另一方面更包括，提出群集分析對於迴歸模型的分群是否適當、計算何種模型距離較為合理、模型的群集數目如何選擇、最終的模型如何判斷，且利用 R 軟體加以記錄計畫的步驟，如此不但可以更有效率的執行計畫，還可以將其成果擴及至其他資料集，然後提供其他研究者，有別於自動選模(向前法、向後法、逐步迴歸法)或僅考慮「單一」準則的另一種選模方式。

2. R 軟體的程式設計：其中又分為三個部分，

(1)收集數據：包括原始數據集的內部資訊、模型的超數據，會利用 R 軟體編寫函數加以儲存及記錄，甚至調整成適用於其他資料集的函數。

(2)群集分析：包括利用 R 套件取得模型的距離矩陣(歐基里得距離)、群集分析的成果，主要是將程式碼做分類並包裝，使程式碼內容不至於過於冗長，如此也讓人更容易理解。

(3)成果報表呈現：最後挑選最佳模型的步驟，依然可以使用 R 函數，條列出所選擇最佳模型的相關資訊(例如：該模型的解釋力、該模型選取變數的個數等)，如此可以使人一目瞭然最後的結果。

3. 個人能力的提升：分為兩個方面，

(1) 專業能力方面：對群集分析更進一步的了解、對報告撰寫的方法論有完整的認識與經驗、對 R 軟體程式設計的技術更純熟。

(2) 軟實力方面：對國外文獻更有方法的研讀、對檢索資訊更有技巧、對尋求師長協助的溝通更有經驗。

(六) 參考文獻

[1]、Konishi, S., Kitagawa, G (2008)。*Information Criteria and Statistical Modeling*。

[2]、劉應興 (1997)。《應用線性迴歸模型》。台北:華泰書局。

[3]、Joseph B. Kadane, Nicole A. Lazar (2004)。*Methods and Criteria for Model Selection.*

[4]、Bovas Abraham, Johannes Ledolter (2005)。*Introduction To Regression Modeling*.

[5]、Mallows,C.L. (1973)Some comments on Cp。*Technomet-rics*,15,661-675。

[6]、多變量分析 (2015)。維基百科。

[7]、黃俊英 (1998)。《多變量分析》。台北：華泰書局。

[8]、Hubert Gatignon (2003)。*Statistical analysis of management data：Boston : Kluwer Academic Publishers*。

[9]、黃博文 (2013)。巨量資料生態雲端策略集群分析-以財務績效指標探討。未出版碩士之論文，國立交通大學管理學院(經營管理學程)，新竹縣。

[10]、曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯 (2013)。《資料探勘》。台北:旗標出版。

[11]、Brand,.etc (2003)。*Metadata Demystified*。

[12]、Ron Kohavi (1995)。*A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*。

[13]、交叉驗證 (2015)。維基百科。

[14]、Arthur Asuncion,David Newman (2007)。Retrieved December 25,2014,from world wide web：http://archive.ics.uci.edu/ml/about.html

[15]、Thomas Lumley (2015)。*R Package"leaps"*。

[16]、Martin Maechler,.etc (2015)。*R Package"cluster"*。

[17]、J.Hair (1995)。*Multivariate Data Analysis,4th ed*。

[18]、Andreas Alfons (2012)。*R Package"cvTools"*。

(七) 需要指導教授指導內容

依照研究流程，可能遇到的困難，分為四個部分：

(1) 準備動作：碩博士及其他期刊論文導讀、國外文獻不通其義時的協助講解。

(2) 研究內容：研究方法的觀念修正，並以客觀角度給予評論、

(3) R 程式：R 程式的包裝及應用、修正程式碼的錯誤。

(4) 成果報告：報表的呈現、撰寫報告的技巧。